Published in final edited form as:
Math Biosci. 2009 June ; 219(2): 57–83. doi:10.1016/j.mbs.2009.03.002.

Recent Developments in Parameter Estimation and Structure Identification of Biochemical and Genomic Systems

I-Chun Chou[*] and **Eberhard O. Voit**
Integrative BioSystems Institute and The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, 313 Ferst Drive, Atlanta, GA 30332, USA

Abstract

The organization, regulation and dynamical responses of biological systems are in many cases too complex to allow intuitive predictions and require the support of mathematical modeling for quantitative assessments and a reliable understanding of system functioning. All steps of constructing mathematical models for biological systems are challenging, but arguably the most difficult task among them is the estimation of model parameters and the identification of the structure and regulation of the underlying biological networks. Recent advancements in modern high-throughput techniques have been allowing the generation of time series data that characterize the dynamics of genomic, proteomic, metabolic, and physiological responses and enable us, at least in principle, to tackle estimation and identification tasks using "top-down" or "inverse" approaches. While the rewards of a successful inverse estimation or identification are great, the process of extracting structural and regulatory information is technically difficult. The challenges can generally be categorized into four areas, namely, issues related to the data, the model, the mathematical structure of the system, and the optimization and support algorithms.

Many recent articles have addressed inverse problems within the modeling framework of *Biochemical Systems Theory* (BST). BST was chosen for these tasks because of its unique structural flexibility and the fact that the structure and regulation of a biological system are mapped essentially one-to-one onto the parameters of the describing model. The proposed methods mainly focused on various optimization algorithms, but also on support techniques, including methods for circumventing the time consuming numerical integration of systems of differential equations, smoothing overly noisy data, estimating slopes of time series, reducing the complexity of the inference task, and constraining the parameter search space. Other methods targeted issues of data preprocessing, detection and amelioration of model redundancy, and model-free or model-based structure identification.

The total number of proposed methods and their applications has by now exceeded one hundred, which makes it difficult for the newcomer, as well as the expert, to gain a comprehensive overview of available algorithmic options and limitations. To facilitate the entry into the field of inverse modeling within BST and related modeling areas, the article presented here reviews the field and proposes an operational "work-flow" that guides the user through the estimation process, identifies possibly problematic steps, and suggests corresponding solutions based on the specific characteristics of the various available algorithms. The article concludes with a discussion of the present state of the art and with a description of open questions.

*Corresponding Author: E-mail addresses: E-mail: bigjump@gatech.edu (I-C. Chou), E-mail: Eberhard.Voit@bme.gatech.edu (E. O. Voit).

Keywords

Parameter estimation; network identification; inverse modeling; Biochemical Systems Theory

1 Introduction

The task of biomathematical modeling comprises the conversion of a biological system into a simplified analogue that is easier to analyze, interrogate, predict, extrapolate, manipulate, and optimize than the biological system itself. The typical approach to mathematical model construction consists of nine phases: (1) data selection; (2) collection of information on network structure and regulation; (3) specification of assumptions and simplifications; (4) selection of a mathematical modeling framework; (5) estimation of parameter values; (6) model diagnostics; (7) model validation; (8) model refinements; and(9) model application (see Figure 1).

The first phase requires the identification and selection of data that are available and suitable to support the purposes of the modeling effort. The second phase is dedicated to collecting information regarding the structure and regulation of the system from the literature and, where feasible, from *de novo* experiments. This phase is confounded by the fact that the true topology of the biological network is often not fully understood and that regulatory details are in many cases incomplete, obscure, or entirely missing. Under these circumstances the task of this and the next phases includes inferences about the network structure and its regulation from biological data. After collecting all relevant information regarding the biological system, the third phase is dedicated to combining this information with additional, acceptable assumptions and simplifications that aim to fill the gaps in the available information. During this phase one also decides which components and interactions of the system should be included in the model. The results are usually visualized as diagrams with nodes denoting the components and arrows representing interactions between them. The fourth phase includes the choice of a suitable mathematical modeling framework and the formulation of symbolic equations. The process usually starts with converting the “wiring diagram” or the “network topology” obtained from the third phase into model equations. In many biological systems analyses, these form a set of ordinary differential equations that represent the dynamic changes in system variables and are governed by fluxes between them. The particular symbolic format of the fluxes depends on the chosen mathematical modeling framework and defines what questions can be asked of the model and what types of methods will be applicable. After the symbolic equations are formulated, the task of the fifth phase is to determine appropriate numerical parameter values that convert the symbolic model into a numerical model and makes the latter consistent with experimental observations. Once this parameterized, initial model is obtained, the sixth phase is dedicated to diagnostics of the model, including analyses associated with steady states, sensitivities, gains, and stability, as well as dynamic features, such as bifurcations and oscillations. In the seventh phase the validity of the model is tested, either with experimental data that had not been used for model construction in a process called cross-validation, or with information from a different biological level. For instance, a metabolic model could be tested against physiological or clinical observations. As presented so far, the modeling process appears to be quite straightforward. However, in most cases it is not linear but cyclic, requiring the return to earlier phases. Addressing the iterative nature of modeling, an eighth phase of model refinement is almost always necessary. Finally, once the model is deemed reliable and appropriate for the stated purposes, phase nine allows the modeler to reap the fruit of the laborious model design. It is now possible to make predictions, generate new, testable hypotheses, suggest the design of novel biological experiments, or manipulate and optimize the model, for instance, toward increases in yield of some organic compound in metabolic engineering, or toward the development of drug treatments in disease.

Among the nine modeling phases, the most challenging task is usually the estimation of parameter values. This estimation is not an isolated task but closely related to other phases in the modeling process. For instance, the size and complexity of the hypothesized model in the second phase have a direct impact on the difficulty of parameter estimation and also affect later analyses as well as the interpretation of results. Most importantly, the choice of the modeling framework naturally influences the degree of complexity, feasibility, and practicability of the parameter estimation task. As a simple example, an explicit linear model permits the use of linear regression methods, which are very well worked out. As soon as the model becomes nonlinear, many of these methods become inapplicable.

Because of the importance of issues related to model selection and to implications for parameter estimation, we will use Section 2 to review the rationale and special demands on mathematical models for biological pathways and to introduce some of the most prevalent and representative modeling frameworks. Generally, model selection and parameter estimation depend on knowledge about the system, the purpose of the modeling effort, and available data. If much is known about the details of the mechanisms governing the biological system, mechanistic formulations, maybe based on principles of physics, are often a good choice. By contrast, if details are lacking, it has been shown that canonical models, such as Lotka-Volterra models and models within Biochemical Systems Theory (BST) are very advantageous for the purposes of mathematical modeling in biological systems. Pertinent details of canonical models will be reviewed in Section 2.3.3.

The development of parameter estimation methods is driven by the availability of experimental data. Different types of data require distinctly different classes of methods. Conversely, the various estimation methods require different types of data. As a pertinent example, data for metabolic pathway models have traditionally characterized the kinetic properties of individual enzymes catalyzing particular steps within a metabolic pathway of interest. The generation of this information occurred hand in hand with the concepts and terminology of enzyme kinetics and the data were measured specifically to parameterize models in familiar formats such as Michaelis-Menten or Hill rate laws. The strategy of using these types of “local” descriptions of model components (one enzymatic process at a time) and merging them into a much more comprehensive, dynamical pathway model is referred to as “bottom-up” or “forward” modeling and will be discussed in Section 3.

Steady-state data may also be used in parameter estimation. These data characterize a metabolic system under conditions where all concentrations have reached constant levels and all fluxes are in balance. Specifically, this type of steady-state analysis is either based on stoichiometric analysis or on experiments that measure the responses of a biological system after a small perturbation around the steady state. Some of these methods will also be discussed in Section 3.

Recent innovations in biological technology enable us to tackle the parameter estimation task in an entirely different, more comprehensive manner, using a “top-down” or “inverse” approach. The biological tools for these purposes are geared toward generating time series data or “global” data of genes, proteins, or metabolites, sometimes under different sets of conditions, such as initial concentrations, or upon various gene knock-outs or the inhibition of specific enzymatic steps. Inverse methods are very appealing, because they provide measurements on cellular or organismal systems in a larger context. In particular, if the data are generated *in vivo* and with only minimal disturbance of the biological system, the insights gained are considered to be as close to reality as is presently possible and potentially much more valuable than data obtained from experiments performed in an artificial *in vitro* setting. Details and features of the traditional and the newly developed techniques with respect to parameter estimation will also be addressed in Section 3.

The potentially high rewards of the inverse modeling approach have motivated scientists from different backgrounds and from all over the world to dedicate considerable effort to the many challenges that must be overcome to make the approach useful. For BST models alone, about one hundred articles and numerous proceedings and book chapters describing computational methods for inverse tasks have appeared within the past decade. We address the specific challenges and requirements of inverse modeling in Section 3.4, along with different types of support algorithms. Many of the pertinent methods target the main problem of optimizing parameter values against the observed time series data. Others suggest strategies for circumventing the costly integration of differential equations, smoothing noisy data, estimating slopes, constraining the parameter search space, or reducing the complexity of the inference task. These auxiliary methods and algorithms will be reviewed in Section 4. The primary focus will be on methods applicable to models within BST, but we will also discuss related issues that are of interest to other modeling approaches.

The earlier discussion of the second modeling phase (see Figure 1) mentioned that the true topology of a biological system is sometimes not fully understood or obscure. In such a case, parameter estimation is much more complicated, because it is *a priori* not clear how to formulate an ill-characterized biological system mathematically. As we will discuss, the use of concept maps [1] and of canonical models is of great help in this situation, because it converts the task of identifying the uncertain structure and regulation of the biological system into a parameter estimation task. Generally, structure identification tasks are much more difficult than parameter estimation, which is already very challenging. Canonical models render both tasks reachable. In particular, one should note that there is no clear boundary between parameter estimation and structure identification if canonical models are used. Section 5 will introduce some of the most relevant structure identification methods, namely the determination of the Jacobian matrix, direct observations, correlation-based approaches, ‘simple-to-general’ and ‘general-to-specific’ modeling, and specially tailored time series data analysis within the framework of BST.

Among all parameter estimation or structure identification methods proposed so far, no single method has risen to the top and can be declared the clear winner in terms of efficiency, robustness and reliability for the majority of realistic case studies. However, it seems that a combination of methods may be useful in a large number of cases. To make inverse modeling more effective and translucent, we propose in Section 6 an operational “work-flow” that guides the user through the various steps of the estimation process, identifies possible problem areas, and suggests promising solutions based on the specific characteristics of the various available algorithms. An interesting consequence, and actually an advantage, of the combined approach is the general result that the optimized solution often consists of multiple, distinctly different parameter sets that are all consistent with the data and that can lead to novel hypotheses for further theoretical and experimental investigation.

Biological systems consist of many organizational layers including genetic, transcriptomic, proteomic, and metabolic levels, as well as phenomena of cell physiology, cell communication, tissue and organ function, populations, and ecology. In this review, we focus primarily on model construction at the genomic and metabolic levels, although many of the computational methods are independent of a particular application. The main reason is that the genomic and metabolic levels are currently best supported by available quantitative data. Metabolic time profiles are particularly well suited because of the stoichiometric property of metabolic pathways, which creates natural constraints on possible parameter combinations, and because its main drivers, namely metabolic concentrations and fluxes, can be measured, at least in principle. In contrast, no material flow is involved in gene-gene, gene-protein, or protein-protein interactions, and the measurable effects are seen in their consequences rather than in characteristics of the processes themselves. Therefore, gene and protein networks must often

be studied with coarser methods than metabolic systems, such as graph theory and Boolean or Bayesian methods, which are applied under the simplifying assumption of binary on/off states. Nevertheless, recent methodological developments have enabled the generation of some dynamic profiles of gene networks, and these have been used for the quantitative identification of gene regulatory networks, primarily by several Japanese groups (see Section 2.3.4). Dynamic data on protein levels are still very rare, and quantitative time series responses are very difficult to obtain. Commensurate with the availability of data, we will primarily focus on the construction and analysis of metabolic pathway models but also discuss issues related to gene interaction networks.

Because the material in this review discusses numerous complementary aspects of parameter estimation and structure identification, it seems useful to summarize the structure of the review in the form of a roadmap, which is given in Table 1.

2 Modeling approaches

2.1 Model requirements

Mathematical modeling and control theory have a long history in physics and engineering. However, the demands and specific requirements in modeling biological systems are quite different and necessitate the adaptation and extension of former methods, as well as the development of novel, additional tools that are optimally suited for modeling biological phenomena. The peculiarities of biological system modeling can be generally described by five aspects (*e.g.*, [2]). First, the biological processes and interactions are highly nonlinear and complex. Thus, a mathematical structure is needed that can capture nonlinearities and does not *a priori* exclude relevant biological phenomena, such as stable oscillations. Second, dynamic responses of biological systems are particularly interesting. Therefore, a suitable mathematical model will have to be time dependent, which almost always requires formulation as a set of differential equations. Third, real biological systems are usually composed of different levels of components and interactions with relatively large numbers, which require the ability of a mathematical framework to scale up to increasingly larger biological models. Fourth, biological systems may have stochastic features when there are only few molecules involved. Under this condition, the fundamental laws of kinetics and thermodynamics are no longer directly applicable and the biological behavior becomes difficult to predict. Thus, in addition to grasping a deterministic phenomenon, the mathematical model should also be able to capture stochastic behaviors when these dominate the process. And fifth, biological reactions rarely happen in a homogeneous environment but are often restricted to surfaces, channels, organelles or compartments. This feature is sometimes important, and therefore the ability of handling spatial processes is necessary for a comprehensive mathematical analysis. It might be added to this list that many biological phenomena evolve over distinctly different time scales and are controlled from different levels of organization. Furthermore, continuous trends are often affected by discrete events, such as the sudden activation of a gene, and by events that lie in the past and cause a delayed effect [3]. At this point, no theoretical or computational frameworks exist to deal with all these aspects. For specific cases, *ad hoc* models may be developed in general purpose software like MatLab or Mathematica, or one might use hybrid methods [4,5], or agent based approaches [6], which however are computationally expensive.

While stochastic, spatial, and time scale effects are known to exist in biological systems, it is often possible to use approximations that greatly simplify model design and analysis. The validity of such approximations needs to be determined on a case-by-case basis. If the approximations are indeed valid, they often lead to simplified system representations based on ordinary differential equations. The generic format for such a representation may be written as

$$\dot{X}_i = V_i^+ - V_i^- = V_i^+(X_1,\ldots,X_n) - V_i^-(X_1,\ldots,X_n),\ i=1,\ldots,n, \tag{1}$$

where X_i denotes the concentration or amount of a variable or variable pool and n is the total number of time dependent variables in the system. The functions V_i^+ and V_i^- represent reactions or fluxes entering or leaving the quantity X_i. This general framework can be recast in numerous alternative ways and only becomes meaningful and specific when the functions V_i^+ and V_i^- are mathematically defined and parameterized. In the following sections we will briefly review some particularly relevant implementations in the context of metabolic pathway analyses and dynamic models of gene regulatory systems.

2.2 Stoichiometric pathway models

Mathematical models describing metabolic pathways can be constructed with a focus either on stoichiometry or kinetics. The stoichiometric property of a pathway is typically considered time invariant, while kinetic aspects are used to capture the dynamics of a system and are driven by the state of the system and may change rather quickly. The stoichiometry of a pathway system determines the wiring diagram of the pathway and describes which fluxes enter or leave which pool and enforces that no mass is gained or lost in the process. The translation of this topological wiring diagram into a matrix equation is straightforward. The resulting stoichiometric matrix contains positive, negative or zero elements that represent which metabolites are converted into which other metabolites. The sign represents the direction of material flow and indicates whether the reaction increases or decreases the concentration of a given metabolite pool. If a metabolite and a reaction are unrelated, the corresponding element is zero. The value of each element in the matrix indicates the stoichiometric relationship and must be an integer. For instance, if one substrate molecule breaks down into two product molecules, the gain in product is coded as +2.

The stoichiometric matrix **N** is the core of stoichiometric models that show how metabolite concentrations change over time. A differential equation is formulated as

$$\frac{dS}{dt} = \mathbf{N} \cdot v, \tag{2}$$

where S is a vector of metabolite concentrations and v is a vector of fluxes. Detailed descriptions of stoichiometric models can be found in numerous journal articles and books (*e.g.*, [7–10]).

The main application of stoichiometric models is the determination of flux rates. Estimation methods for this purpose depend on the type of available experimental data. In most analyses, stoichiometric models are studied for metabolic systems in a steady state where, for all pools, the material flow into the pool equals the material flow out of the pool and all flux rates are constant. Under this assumption, the left-hand sides of the equations in Eq. (2) become zero and the system of differential equations turns into a set of linear algebraic equations. If the stoichiometric matrix is full rank, it is straightforward to calculate the fluxes. However, usually there are more unknown fluxes than equations, so that the system of linear equations is underdetermined. Stoichiometric analysis is most often applied to microbial systems, where it is assumed that the microbes tend to optimize their growth rate. This assumption can be formulated as an objective to maximize the availability of nutrients needed for growth. Representing this objective with a linear function transforms the underdetermined stoichiometric system into a linear programming task, which is easily solved even for large

systems. Mass conservation and stoichiometry are distinctive properties of metabolic pathways and not applicable to the dynamics of gene regulatory or proteomic networks.

Flux balance analysis (FBA) inherits the properties of the stoichiometric approach but additionally imposes physical and chemical constraints to find the feasible or optimal distribution of fluxes [11]. For instance, it is possible to account for thermodynamic limitations. The background of FBA is reviewed in Palsson [9] and representative developments of variations are summarized in Kauffman *et al*. [12]. The modeling process in FBA consists of system identification, mass balancing, defining measurable fluxes, and optimization. System identification consists of setting up the stoichiometric equation and relevant constraints. Mass balance is the application of stoichiometry and conservation of mass. For instance, the total number of moles of carbon in the system is conserved during the time of reaction. By accounting for all material flows entering and leaving each metabolite pool in the pathway, one can determine the material distribution and also identify flows that might have been unknown or difficult to measure in experiments. As in stoichiometric analysis, the optimization step assumes an objective such as maximization of growth rate or product yield that permits solutions through linear programming. The main advantage of both the stoichiometric model and FBA is the linearity of their matrix representation at the steady state, which tremendously simplifies the analysis, since there are numerous well-established methods of linear algebra that are directly applicable. Several examples have demonstrated that FBA is capable of assessing the theoretical capabilities and operative modes of metabolic systems in the absence of kinetic information (*cf.* [9,13–16]).

Stoichiometric models are sometimes studied under a pseudo-steady-state (PSS) assumption in cases where the concentrations of metabolites rapidly adjust to new levels [17–19]. This PSS approximation was shown to be valid for most intracellular metabolites [20]. Under this assumption, it is reasonable to neglect the instantaneous changes of metabolites and set the rate of change to zero.

When complete time courses of metabolites are available, the flux distribution at each time point can be determined with [20] or without [21] the PSS assumption. Distinctly different from the standard application of stoichiometric analysis, where only steady-state data are used, the dynamic changes in metabolite concentrations in the latter case are not necessary zero and can be used to gain incomparably deeper insight into the pathway at hand [21].

Mahadevan and coworkers [22] extended traditional FBA to account for dynamics and presented two different formulations: the dynamic optimization approach (DOA) and the static optimization approach (SOA). DOA involves optimization over the entire time period of interest to obtain time profiles of fluxes and metabolite levels. SOA involves dividing the batch time into several time intervals and solving the instantaneous optimization problem at the beginning of each time interval. By testing the methods in the analysis of diauxic growth in *Escherichia coli*, the authors concluded that SOA was computationally simpler to implement provided all of the constraints were linear, whereas DOA was more flexible and suitable for the incorporation of experimental data.

Utilization of the stoichiometric property together with dynamic changes in metabolites is a valuable option for studying flux distributions in metabolic pathways. However, the main advantage of the standard stoichiometric approach, namely linearity at the steady state, is at the same time its most severe limitation. Specifically, this approach focuses almost exclusively on the connectivity structure of the system and the flux distribution at steady state, but does not account for kinetic features, which often are necessarily nonlinear. Therefore, the predictive power of linear stoichiometric models, while successful in many cases, is also limited because regulatory signals and other nonlinear dynamic interactions cannot be included in the model

without destroying its linear structure. As a compromise, Palsson and others (*e.g.*, [23]) introduced binary-valued regulatory matrices that are multiplied to the stoichiometric matrix and determine whether a given flux is activated or not. However, a full account of nonlinear regulatory features requires the formulation of a pathway system as a kinetic model.

2.3 Kinetic models of pathway steps

When detailed information is available about the kinetics of the metabolic reactions in the pathway, it is possible to describe its dynamics by incorporating kinetic features in the flux representations v of the general stoichiometric models in Eq. (2) [24]. The crucial step toward combining the stoichiometric property with kinetic features is to search for appropriate functional forms to represent the flux quantities V_i^+ and V_i^- in Eq. (1), which then translate into representations of the vector v. Approaches for this purpose can be categorized into three categories (see Figure 2): (1) mechanistically based functions (*e.g.*, law of mass action, Michaelis-Menten rate law); (2) *ad hoc* approaches; and (3) different types of canonical models (*e.g.*, BST and lin-log representations). Details of these representations are reviewed in the following sections.

2.3.1 Mechanistically based functions

Mass action systems: Models based on the law of mass action are typically used to describe reaction networks consisting of elementary reactions. The rate of a given elementary chemical reaction is proportional to the product of concentrations of all variables reacting in the elementary process, including their moieties, and is generally formulated as the basis function

$$v=k\prod_{g=1}^{n}X_i^{g_i},i=1,2,\ldots,n, \tag{3}$$

where k is the rate constant, which is always positive, and g_i are kinetic orders which are nonnegative integer numbers that reflect the numbers of molecules involved in the reaction. The main advantage of models based on the law of mass action is that they can be determined directly from the elemental reactions and their stoichiometry. The drawback is that most *biochemical* reactions are not elemental but catalyzed by enzymes, and therefore composites of several elementary reactions [25]. It is inconvenient and indeed infeasible to carry all these reactions along and much easier instead to develop composite rate functions. Furthermore, the underlying mechanisms of an enzyme catalyzed reaction are not always understood in sufficient detail or they are experimentally inaccessible. Therefore, mass action equations are difficult to set up and parameterize for complex pathway systems.

Michaelis-Menten and similar rate laws: The Michaelis-Menten rate law (MMRL) [26] and its variations are among the most commonly used representations for kinetic modeling in metabolic pathway analysis. MMRL is based on the concept that a substrate and an enzyme form a transient complex which either dissolves to return the two or leads to the formation of a product and the release of the enzyme. The modeling of enzyme reactions in this approach is simplified considerably under the quasi-steady-state assumption, which states that the intermediate complex does not change appreciably over time. Even though Michaelis-Menten rate laws are often straightforward to set up, complete descriptions of more complex enzyme mechanisms may become massive if several substrates or reactions are involved [25]. As the result, mathematical analyses become very complex and the parameter estimation requires an undue amount of experimental data [27,28]. In addition to technical issues, the model results are difficult to interpret in terms of the underlying biological system [28,29]. The estimation

of parameter values in pure MMRL is easy and may even be accomplished with methods of linear regression [30]. However, this simplicity vanishes for larger systems of MMRLs and their generalizations.

2.3.2 *Ad hoc* modeling approaches—When the detailed mechanisms of a biological process are unknown or unclear, an alternative approach is to adapt a mathematical model from a different context, quasi as a black box model, to describe the biological phenomenon of interest. As a typical example, Voit and Sands [31] reviewed a collection of mathematical models for the uptake of nutrients by tree roots, most of which had no foundation in plant physiology but were expected to fit observed trends sufficiently well. Although such an *ad hoc* black-box approach might provide models that fit the observed data, it is highly arbitrary and poses several intrinsic problems. For instance, one must question the validity of such a model for other experimental data or extrapolations to un-tested experimental conditions and anticipate problems with interpretations of results, comparisons of results with those from alternative models, and extensions to more refined models. Furthermore, the estimation of model parameters may become problematic, because they do not necessarily correspond to measurable quantities. In many cases, a better option than an *ad hoc* model is a canonical model.

2.3.3 Canonical models—The predictive ability of stoichiometric models is limited because nonlinearities due to regulation cannot directly be accounted for. For improved models that permit nonlinearities, kinetic information of the pathway is needed. A good compromise that is capable of capturing nonlinear dynamics while keeping the mathematics relatively simple is the use of a "canonical" nonlinear model whose structure is fixed and whose individuality comes from its parameter values. In addition to their homogeneous structures, canonical models are more or less size independent and facilitate the development of customized techniques and methods for analysis, diagnostics, and parameter estimation.

Arguably the most promising canonical nonlinear models in metabolic modeling are Generalized Mass Action (GMA) and S-system structures within *Biochemical Systems Theory* (BST) [27,32–36]. These models are constructed by approximating fluxes with products of power-law functions, which are mathematically grounded in the well-established approximation theory of Taylor. In the S-system formalism, each equation has a particularly simple format: The change in system variables is given as one set of influxes minus one set of effluxes (*cf.* V_i^+ and V_i^- in Eq. (1)), and each set is collectively written as one product of power-law functions. Thus, the generic S-system formulation reads

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}, i=1,2,\ldots,n, \tag{4}$$

where X_i represents a time-dependent variable (metabolite) and n denotes the number of variables in the system. The non-negative multipliers α_i and β_i are *rate constants* which quantify the turnover rate of the production or degradation, respectively. The real numbers g_{ij} and h_{ij} are *kinetic orders* that reflect the strengths of the effects that the corresponding variables X_j have on a given flux term. A positive value signifies an activating or augmenting effect exerted by X_j, a negative value signifies an inhibitory effect. A kinetic order of zero implies that the corresponding variable X_j does not have any effect on a given flux. In some instances, m independent variables, which are typically constant during each mathematical experiment, may be included. They do not have their own equations but enter the power-law terms just like dependent variables, so that the products run from 1 to $n+m$.

In the GMA formalism, instead of aggregating all influxes and all effluxes into one term each, all influxes and effluxes are approximated individually with power-law terms such that

$$\dot{X}_i = \sum_{k=1}^{k_i} \left(\pm \gamma_{ik} \prod_{j=1}^{n} X_j^{f_{ikj}} \right), i = 1,2,\ldots,n, \tag{5}$$

where the rate constants γ_{ip} are non-negative and the kinetic orders f_{ipj} may have any real values as in the S-system form. Also as before, independent variables may be included. It should be noted that differences between these two formulations only exist at branch points, whereas all other steps are identical. One should also note that S-systems permit parameter estimation methods [37], which are not directly applicable to GMA systems, as we will discuss later.

BST models have a number of important advantages which have been discussed in detail elsewhere [27,32,33,35,36]. Four are particularly crucial here. First, these systems are rich enough in structure to capture virtually any nonlinearity including complex oscillations and chaos [38,39]. Second, symbolic BST models can be set up without mechanistic information on the underlying system, but if information is available, it can be used to simplify the symbolic representation. Third, the highly structured format facilitates mathematical and numerical analyses. These analyses include computations associated with steady states, sensitivity, stability, as well as dynamic features. Fourth, BST models are characterized by a one-to-one relationship between parameters and structural features. Thus, if structural features are known, it is explicitly clear where they will appear in the BST models. Conversely, if a parameter has been identified, its interpretation in terms of structural properties is immediate. This feature is especially crucial for structure identification and parameter estimation of metabolic models. The reasons will be discussed in detail in Section 4.4. The power-law models of BST were initially used to model metabolic pathways and gene circuits, but this formalism has also proven very beneficial for modeling other classes of biological systems, including genetic networks, immune networks, multi-level systems, and cell signaling [40–45].

An alternative canonical form is the “lin-log approximation” which was introduced by Hatzimanikatis and Bailey [46] and expanded by Visser and Heijnen [47]. This form is based on taking the logarithm of each metabolite concentration and enzyme activity in relationship to a corresponding reference value. The lin-log model constitutes an extension of Metabolic Control Analysis (MCA), a theoretical framework for analyzing control and regulation in metabolic networks close to their steady state [29,48,49]. For small variations about a steady state, BST and lin-log models show similar responses. However, they differ for larger variations [50–52].

Another recently proposed canonical form is the *Saturable and Cooperative Formalism* (SC formalism) [53], which is based on Taylor approximation in a special transformation space that is defined by logarithms of power-inverses. The SC formalism exhibits improved cooperativity and saturation in comparison to other canonical formalisms. In addition, the SC formalism is expected to be accurate over a wider range around the operating point if the approximated functions are saturated. The main drawback of the SC formalism is its need for a much larger number of parameters, which on one hand bestow SC models with higher flexibility but on the other hand require increased estimation efforts.

For completeness, we should mention Lotka-Volterra (LV) models and their generalizations [54–57]. These models are canonical in the sense of the models discussed before and have found very many applications in ecology. In fact, they are very flexible in structure and, with

sufficiently many variables, can model any nonlinearities that can be expressed as ordinary differential equations, just like BST models [38,58,59].

The generic LV format is

$$\dot{X}_i = X_i \cdot \left(a_{i0} + \sum_{j=1}^{n} a_{ij} X_j \right) \tag{6}$$

where the coefficients a_{ij} are real numbers. In other words, the i^{th} equation contains one term that is linear in the variable itself and a product of the variable with another system variable. While this format is very useful in ecological systems analysis, it is less so in metabolic systems. For instance, it does not allow the direct formulation of an un-modulated precursor-product relationship, because the i^{th} equation cannot contain a term of the form kX_{i-1}.

The choice of an S-system, GMA, lin-log, or SC model depends on the available input information and the purposes of the intended model. For instance, GMA systems are natural extensions of stoichiometric models that incorporate kinetic information using a power-law approximation and are closer to biochemical intuition than S-systems. However, the GMA format, as well as the SC format, does not allow the algebraic calculation of steady states, which is important for a number of analyses. The lin-log model shares the intuitive advantage with the GMA format and, like the S-system format, allows algebraic steady-state calculations. However, it cannot represent certain nonlinear behaviors since its structure is essentially linear [60]. Also, the lin-log approximation results in substantial errors for substrate values close to zero [50,52] and may lead to negative rates if the substrate concentration is low, whereas the BST representations become more inaccurate for very high substrate concentrations. As local approximations, all these formats perform similarly as long as the system variables do not deviate too much from some chosen state where they coincide.

2.3.4 Dynamic models of gene regulatory networks—The previous sections reviewed representative modeling strategies for metabolic pathway systems and suggested that power-law models are particularly useful in this context. The same power-law formalism is also beneficial for modeling gene regulatory systems, although there are other distinct alternatives (for a review, see *e.g.* [61]). Mathematically the simplest representations are linear; they typically capture the structure of the network with a connectivity graph and its associated connectivity matrix (*e.g.*, [61–63]). The graph is modeled as static, that is, as time invariant, and allows classifications of the particular type of the network (*e.g.*, "randomly connected," "hub and spoke," or "small world"), as well as analyses of the robustness to different types of perturbations (*e.g.*, [64]). A different modeling strategy is based on Boolean networks, in which each gene is either turned on or off, and whose discrete dynamics is governed by rules that recursively determine the expression state of a gene during the transition from one time point to the next (*e.g.*, [65]). Gene networks may also be analyzed with Bayesian methods that assess the probability of a gene being affected by other genes (*e.g.*, [66,67]).

Of particular interest here are models that trace the full dynamics of gene networks with ODE models. While many alternatives could be used as base models, many researchers have resorted to the default of S-system models, because these do not require knowledge of the precise mechanisms that govern the regulation network. The study of gene regulatory networks with S-systems has a long history, which began with the pioneering work of Savageau [68]. Scientists had been observing many types of gene circuitry and Savageau and his collaborators tried to answer questions such as: Is each type of circuitry organized in a manner that is optimized within its context or is the mode of regulation a coincidence? Can gene regulatory

circuits be classified into types that are predictable from the organism's surroundings (*e.g*. activation *versus* repression, positive *versus* negative control)? Is it possible to deduce the functionality of a given circuitry from its particular organization? Is it possible to infer general "design principles" for the modes of regulation from the analysis of alternative gene circuitries? Is it possible to implement these principles and "synthesize" gene circuits *de novo* in order to achieve prescribed biological responses (*cf.* [40,69,70])? To answer these questions, Savageau and others dedicated significant effort to studying gene regulatory circuits with techniques of canonical modeling and the method of controlled mathematical comparisons (*e.g*. [43–45, 71–74]). A particularly intriguing result was the conceptual framework of *demand theory* (*e.g*. [75–80]), which indeed tied types of gene regulation to environmental demands and thus revealed natural design principles.

One of the models used to analyze gene circuits [72] has become a benchmark system for inverse methods (see Table 3). In addition, Maki *et al.* [81] proposed a network of 30 genes (see also Kimura *et al*. [82] and Kutalik *et al*. [83]), which is artificial but was inspired by realistic gene networks, for testing the performance of algorithms on larger systems. In a similar vein, Kutalik *et al.* [83] created an artificial network of 7 genes.

3 Kinetic model construction

The collection of input information and the choice of a mathematical model framework result in a symbolic model which is typically in the form of ordinary equations, as discussed before. The next step is to assign numerical values to all parameters in the model. There is no unique recipe for this task of parameter estimation. In fact, the estimation problem is almost always complicated and continues to be the bottleneck of biomathematical modeling.

In this section, we review the classes of parameter estimation methods that are in current use, namely, forward (bottom-up) approaches, estimation from steady-state data, and inverse (top-down) modeling using time-series data. The nature of suitable data for each type of estimation is distinctly different, and so are the methods of analysis. Ideally, rich and diverse data will allow the use of several methods that complement each other or a combined strategy that has much greater potential leading to suitable models than either approach by itself [1,84–86].

3.1 Forward or bottom-up modeling

Before high-throughput data were available, essentially all metabolic models were developed from "local" kinetic information of biochemical or physiological responses, which had been obtained in the traditional reductionist manner. Typically, biologists around the world worked on characterizing one particular enzyme or transport step at a time. They purified the enzyme, studied its characteristics, determined optimal temperature and pH ranges, and quantified cofactors, modulators, and secondary substrates. Isolated from these laboratory experiments, modelers converted this information into alleged mathematical rate laws. Once enough information had been collected for most steps in the pathway, the modeler attempted to merge this information into an integrative mathematical model. This type of model construction has been benefiting greatly from databases like KEGG [87,88], MetaCyc [89], and Brenda [90], which contain large collections of pathway topologies and kinetic parameters retrieved from literature sources. In many cases, the integration into a viable model does not succeed at first and requires revisiting assumptions and parameters, or even the search for additional experimental data.

If done right, this "forward" or "bottom-up" process leads to a model representation of the pathway that exhibits the same features as reality, at least qualitatively, if not quantitatively. Case studies that used this forward estimation approach in BST successfully include: the TCA cycle in *Dictyostelium discoideum* [91]; the citric acid cycle [92,93]; fermentation in

Saccharomyces cerevisiae [94–96]; purine metabolism [97–99]; the Maillard-glyoxylase network with formation of advanced glycation end products [100]; the ferredoxin system with information from protein structure for model identification [101]; and sphingolipid metabolism in *Saccharomyces cerevisiae* [102–104]. In almost all of these cases, the strategy consisted of setting up a symbolic model, estimating local parameters, studying the integration of all individual rate laws into a comprehensive model, testing the model, and making refinements to some of the model structure and the parameter values in an iterative fashion.

While theoretically straightforward, there are several disadvantages of this approach. The main issue is that a considerable amount of local kinetic information is needed and that this information is often only available from different organisms, different species, or experimental performed under different conditions. As a consequence, more often than not the "integrated result" is not consistent with biological observations. Furthermore, this process of construction and repeated refinement is very labor intensive and requires a combination of biological and computational expertise that is still rare.

3.2 Model retrieval from steady-state data

By far the most model estimations from steady-state data have been performed in the context of stoichiometric and flux balance analysis, as discussed before. Specifically, fluxes entering and leaving a metabolic system were measured under the assumption of complete metabolite and flux balancing, and internal flux rates were inferred from the assumed stoichiometric models and the maximization of growth rate per linear algebra and linear programming. Examples are plentiful (*cf.* [9,10]).

Using concepts of stoichiometry in a slightly different fashion, Wiechert and others developed isotopomer and cumomer methods for flux rate estimation using radioactive labeling techniques [105–108]. The technique is based on the fact that specifically labeled atoms in an input molecule, such as 1-^{13}C-glucose, have known metabolic fates. By letting the system resume steady state after the labeled input, the distribution and positioning of label among the metabolites of the entire pathway provides clues of internal flux rates. The original steady-state estimation methods based on positional labeling were later extended to characterize the transients between input and steady state [109,110].

A distinctly different type of parameter estimation from steady-state data uses responses of a biochemical system to (infinitesimally) small perturbations around the steady state. Two types of approaches may be taken. First, parameter values can be obtained, at least in principle, by direct experimental measurements of how a variable affects the fluxes entering and leaving the metabolite pool [33,48,111]. Suppose the flux rate and metabolite concentrations in steady state of one particular biochemical process are known. One can then slightly alter the concentration of a variable systematically while keeping the other variables constant. The result of these experiments can be plotted as flux rate versus metabolite concentrations in logarithmic coordinates [33]. In the case of power-law systems, the kinetic order of the investigated variable corresponds to the slope of the plotted line and is obtained by linear regression (*e.g.*, [35, 103]). Under ideal circumstances, sufficient experimental measurements can be collected to allow the regression analysis. However, the data usually contain noise and consist of only a few measurements, which make the regression rather susceptible to experimental uncertainties. As an alternative, parameter values can be estimated by experimental measurements of logarithmic gains (*e.g.*, [111,112]). This approach is based on perturbing variables in the interesting portion of the pathway and recording the corresponding changes after perturbations. The information about modulation, including flux rates and concentrations, is collected to calculate the kinetic orders. Additional methods that have some similarity to these steady-state estimations are discussed in Section 5.1.1.

3.3 Inverse or top-down modeling

While steady-state measurements or simple perturbation experiments around the steady-state can effectively be used for an estimation of flux rates, incomparably more information about the system is contained in measurements of metabolite concentrations at sequential points in time that may include wider deviations from the steady state. Modern high-throughput techniques of biology are capable of producing these types of time series data, and they have begun to render distinctly different options for modelling metabolic systems possible. Because this type of estimation begins with comprehensive data at the level of intact systems and leads to inferences of parameter values at lower levels of organization, namely features of individual processes, it is called "top-down" or "inverse." The primary experimental tools for measuring genomic time series data are microarrays and RNA-based gene silencing (*e.g*. [61,113]). For proteomic time series, two-dimensional gel electrophoresis and mass spectrometry could be employed, but technical challenges have kept the number of proteomic time series analyses small (*e.g*. [114]). Dynamic metabolite concentration profiles are presently obtained with methods of nuclear magnetic resonance (NMR) (*e.g*. [115,116]), mass spectrometry (MS) (*e.g*. [117,118]), and high performance liquid chromatography (HPLC) (*e.g*. [119,120]). In contrast to the "local" data obtained from traditional experiments, the clear advantages of using "global" data are that the information is collected within the same organism, obtained under the same experimental condition, and sometimes even *in vivo*. Time series data contain enormous information on the structure and regulation of the biological system they describe. However, this information is mostly implicit, and it is very challenging to extract it from these data due to the complexity and nonlinearity of biological networks. There are several distinct challenges of this approach, some of which are readily anticipated, while others are surprising and puzzling. We describe these challenges in detail in the next section.

3.4 Challenges of the top-down modeling approach and current solution strategies

The challenges of model identification from time series data may be the result of biological and/or technical features of the case. They generally fall into four categories, namely: data related issues, model related issues, computational issues, and mathematical issues [85,121, 122]. Responding to these challenges, numerous modeling techniques for the analysis of dynamic data have been proposed. Representative examples are summarized in Table 2 and described below and in later sections.

3.4.1 Data related issues—Biological datasets usually contain noise and measurement errors, and they are seldom complete. Typical scenarios of missing data points include that the data are sparsely missing, that data collection is lacking at certain time points, or that entire time series are missing, either due to technical issues (*e.g*. the concentration is below the detection limit) or to the possibility that relevant metabolites may be unknown. In simple cases of relatively few missing points, standard interpolation and smoothing methods may lead to satisfactory solutions. Many methods are available to address these tasks; some will be discussed in a later section.

In other cases, data may be missing for entire time courses, yet qualitative or semi-quantitative information is available. To bridge this type of gap between specific wet experiments, biological insight and intuition, and the construction of mathematical models, one may pursue the strategy of concept map modeling, which permits the inclusion of semi-quantitative information on expected responses of the system based on biological knowledge and intuition [1]. We will further describe the general features of concept map modeling in Section 7.

Besides missing data points or time series, other data related problems must be addressed. For instance, a pathway model may be designed in such a fashion that all mass is accounted for throughout the experiment. However, if the experimental data are noisy or incomplete, it is

possible that non-negligible amounts of mass are apparently lost or even gained. This inconsistency may cause problems for the estimation of parameter values. We discuss this issue in more detail in Section 6.2.2.

Even if the time series are complete, they are almost always noisy. They are also often affected by uncertainties about the particular experimental conditions at the time of observation. For instance, temperature and pH may affect the reaction mechanism or speed, but they are not always reported. As far as possible, uncertainties should be taken into account as these can possibly influence the parameter values and thereby have an impact on the predictive accuracy of the resulting model.

Other potential problems in the dataset may include that the time series are non-informative, *e.g.*, consist essentially of constant time profiles, or that they are collinear, as is the case when some of the variables are proportional to each other along the entire time horizon. This collinearity between time series might cause ill-conditioning in the estimation process. The problem should be diagnosed beforehand by checking the conditional number or correlation coefficient of the data, and remedied if possible by pooling variables or merging essentially constant variables with the rate constants of the term. We will return to this issue in Section 7.

At the opposite side of the spectrum of data availability, it may also happen that data have been measured but that they are difficult to include in the model. This situation occurs especially with "ubiquitous" metabolites like ATP and water, which may clearly affect the dynamics of a metabolic pathway system, but are involved in so many processes that a comprehensive model simply cannot be constructed. A possible solution is a "partial modeling" strategy that permits the mixing of well-defined components with components whose dynamics is only known in the form of time series that are observed but cannot be formulated in terms of other model components (*e.g.*, [1,122]; see [82,123,124] for a similar strategy in a different context). As a specific example, we analyzed time series data describing glycolysis and lactate production in *Lactococcus lactis* [121]. The data contained ATP and NAD^+ concentrations, measured over time, but it was impossible to formulate the ATP and NAD^+ dynamics as functions of the system variables, because both are involved in very many reactions, most of which were not modeled. As a solution, the better-defined components were formulated as differential equations in BST, and their dynamics included ATP and NAD^+ as variables. However, ATP and NAD^+ were not modeled as differential equations but as "off-line" data, which entered the system as time-dependent "forcing functions."

3.4.2 Model related issues—The inverse problem requires a mathematical model that captures the dynamics of the data in a suitable fashion. However, there is an unlimited variety of nonlinear structures and mathematical formulations that could be potential candidates for the optimal data representation. We have introduced some of the modeling frameworks and their pros and cons in Section 2. Here we highlight the specific challenges associated with model selection in top-down modeling approaches.

It seems that there are good reasons for selecting model formalisms that are intended to represent the underlying chemical reactions mechanistically. However, this mechanistic approach is not always the best choice. The reasons are, first, that high-throughput time series data are essentially never of sufficient accuracy to discern among possible underlying reaction mechanisms. Second, the true mechanisms underlying the data are simply not known. Third, traditional kinetic rate functions, such as the Michaelis-Menten rate law, are not necessary the best choice for *in vivo* data [125,126]. Instead of trying out a roster of different mechanistic formulations that could potentially be appropriate, it is often more efficient to use a generic nonlinear approach in the form of a suitable approximation that is based on criteria like: the ability to capture important mathematical features of a data set, simplicity of representing the

data, mathematical tractability, and interpretability of mathematical results within the biological realm. Canonical models, as they were described before, are particularly useful for this purpose.

3.4.3 Computational issues—The estimation process itself is challenging computationally, especially if the model consists of nonlinear differential equations [127]. It is no surprise that the difficulties grow with the size and complexity of the system, which usually translate into increasing numbers of equations and variables in the model. None of the nonlinear methods are truly straightforward, and even for systems of modest size, all lead to challenging issues, such as slow algorithmic progress toward the error minimum, complicated error surfaces, lack of convergence, or convergence to local minima. Furthermore, the differential equations need to be integrated during the optimization process, unless special strategies are employed that explicitly avoid this step. The integration may be very time consuming, especially when the system is stiff, and require 95% or more of the entire estimation time [128]. Other computational challenges include the distinction between direct and indirect effects among system variables, characterization of intermediate steps, time delays, spatial heterogeneity, and stochasticity at the level of the governing processes or the integrated system.

The actual development of algorithmic methods for extracting information from biological time series data sets is the subject of Section 4. The methods for biological inverse problems typically require a combination of techniques that include techniques for attacking the main problem of optimizing parameter values, as well as supporting algorithms, such as methods for circumventing the costly integration of differential equations, smoothing overly noisy data, constraining the parameter search space, or reducing the complexity of the inference task.

3.4.4 Mathematical issues—An often ignored source of problems is a (frequently unknown) mathematical redundancy in some models. Redundancy may occur in different manifestations. It is possible that different sets of parameter values, which fit the experimental data exactly equally well, are mathematically or numerically equivalent [129]. For instance, if two parameters p and q always enter the equations in the same combination, such as $(p+q)$, it is not possible to identify their individual values. In the context of power-law functions, collinearity between variables leads to unidentifiability of the corresponding kinetic orders [130,131]. It may also happen that solutions exhibit similar residual errors, even if they are mathematically not equivalent. One possible cause is compensation between a rate constant and the kinetic orders in the power-law terms of a particular data fit [130,132]. Mathematical redundancies and error compensation may occur within or between flux descriptions in the same or among different equations. The task of dealing with mathematical and computational redundancies has been addressed in some articles [130,132,133], but will require more work. A preliminary step in the identification of redundancies and their causes may be our recent method of Dynamic Flux Estimation [21], which permits unique tools of error diagnostics (see Section 7).

Despite these challenges, inverse approaches based on *in vivo* time series data are certainly worth pursuing, because these data are the most accurate reflections of what cells and organisms really do in environments that are as close to reality as is presently possible to measure. This level of realism is very appealing, and many researchers have worked on the development of methods that overcome some of these challenges. A representative set of methods is described in the following section.

4 Parameter estimation techniques for top-down modeling approaches

Many of the recently developed techniques for top-down parameter estimation have been developed for BST models. Most of them are similarly applicable to other canonical models,

although a few take advantage of the specific form of power laws in BST. The main algorithms and their representative references are shown in Figures 3 and 4. A historic listing of representative algorithms is presented in Table 3.

4.1 Methods based on integrating differential equations

The central component of solving inverse problems is an efficient algorithm for determining optimal estimates. Most standard methods for this purpose naturally involve the numerical solving of the differential equations. This solution is computationally very expensive. As a specific example indicative of the problem at hand, consider a direct attempt to estimate the parameters of a five-variable S-system model from noise-free time series data with a genetic algorithm [134]. The authors used a cluster of 1,040 CPUs, which ran for ~10 hours for each loop of the estimation program. Needing 7 loops, the entire estimation time thus was roughly 70,000 PC-hours. Analyzing this alarming situation, the distinct tasks within the optimization were clocked in detail with the result that 95% of the time spent for parameter searches involving differential equations is used for integrating the equations, while relatively little time is used to compute gradients toward the optimal estimates [128]. If the equations are stiff, the computation time may increase to almost 100%, and even if the model is not stiff, the likelihood is high that some trial solutions during the algorithmic process could make it stiff [128]. Therefore, even if efficient, custom-tailored integration methods are used [135], significant time savings can be gained by speeding up the evaluation of differential equations.

Numerical integration of the ODE system can be circumvented when the differentials are substituted with slopes that are estimated from the time series data at all measured time points. This substitution entirely eliminates the need to integrate differential equations, because the parameter estimation is subsequently executed on systems of algebraic equations. Furthermore, the equations become uncoupled so that they can be assessed in parallel or one at a time. An early implementation of this method was accomplished by manually estimating slopes from observed time series data and substituting them for the derivatives in the differential equations [35,136,137]. Voit and Almeida [128] implemented the slope-estimation-decoupling strategy with an artificial neural network that simultaneously permitted data smoothing and the computer-algebraic computation of slopes from the smoothed time courses.

The slope-estimation-decoupling idea has subsequently been combined with various methods such as genetic algorithms, simulated annealing, swarm methods, interval analysis, and a number of hybrid methods. Various slope estimation methods will be reviewed in detail in Section 4.2. One drawback of these methods is that it may not be easy to obtain good measurements or estimates of the slopes if the data are very noisy. However, it is still advantageous to use this approach, even if the results are coarse, since the coarse estimates may be used as initial guesses for standard nonlinear optimization methods. Other advantages of the decoupling approach are reviewed in [128].

In a different implementation of a similar idea, the decoupling allowed solving and fitting of one differential equation at a time instead of solving the entire system. Maki *et al.* [123] proposed this “step-by-step” strategy and Kimura *et al*. [82,124] introduced a similar concept called “decomposition,” which dissects a large network inference problem into many smaller sub-problems. In both methods, the variables contributing to the single differential equation being integrated are substituted with the actual observed time series data or with smoothed analogues, which are thus used as off-line inputs to the decoupled system. This approach significantly reduced the computation time. For instance, using the same artificial five-variable datasets that required 70,000 PC hours [134], Kimura and co-workers ran the algorithm on a single CPU, where it required only about 59 minutes to optimize each subtask.

A drawback of decoupling and decomposition approaches is that each subtask is solved independently, a procedure which does not allow the exchange of information between subtasks. For instance, the variables serving as off-line data in one equation are actually solved in another equation. Thus, if the value of one variable is updated during optimization, the information should be incorporated into optimization processes of the other subtask. This feature is especially important when there is considerable noise. Kimura *et al*. [82] proposed to solve the decomposed subtasks simultaneously using a cooperative coevolution algorithm. Since the decomposed subtasks interact with each other through their calculated time series data, the inferred model is more likely to represent the dynamics correctly.

In order to reduce the number of numerical integration steps, Matsubara *et al.* [138] proposed to use a radial basis function network (RBFN) for parameter estimation. RBFN is a type of artificial neural network that uses radial basis functions as activation functions; it has been shown to be able to approximate nonlinear time series data efficiently [139]. In order to examine the performance of RBFN, Matsubara and co-workers proposed two schemes: one used a simple genetic algorithm (SGA) with numerical integration, and the other combined RBFN with a genetic algorithm in the input data selection phase. Both schemes were examined in metabolic pathways using Michaelis-Menten equations. While SGA improves the fitness between parameterized model and time series data and integrates every time during optimization, RBFN predicts the optimal parameter values by learning the relationship between parameters and fitness values using slopes to replace derivatives and integrates the system only once at the last step. Therefore, numerical integrations used to evaluate the fitness are reduced from many to one. The results indicated that the RBFN scheme halved the computation time and increased the success rate of the optimization task.

An alternative approach avoiding numerical integration is a modified collocation method, which converts ordinary differential equations into algebraic equations which directly adopt the measured data to approximate the dynamic profiles at sampling points. This approximation not only reduces computation time, but also decouples the equations so that parallel computation is possible for the parameter estimation. A collocation method was combined with hybrid differential evolution (HDE) to determine the global solution of an estimation task [140]. Again, applying this type of "uncoupling" strategy in combination with other estimating methods reduced the computation time dramatically.

4.2 Slope estimation

As a crucial part of the slope-estimation-decoupling strategy, decent estimates of the slopes are required. If the data are more or less noise-free, simple linear interpolation, splines [141–143], B-splines [144], the so-called three-point method [145], or even hand fitting [137] is effective. If the data are noisy, it is useful to smooth them, because the noise tends to be magnified in the slopes. Established smoothing methods again include splines, as well as different types of filters. Artificial neural networks (ANNs) have been shown to be useful in a number of applications of biochemical pathways modeling [146]. They are so general and flexible that they are considered "universal functions" that are obtained from training the ANN [128,147–149]. The main advantage of using ANNs is that the resulting time traces can be trained to fit the data arbitrarily closely and that they have an algebraic format for which the slope can be computed straightforwardly with methods of computer algebra [146,150]. Furthermore, the universal output function provides an unlimited number of interpolated data points within the time interval of interest. Other advantages of ANN are reviewed in [146] and [128]. The ANN method was shown to determine the smoothed traces very efficiently even if the data contained considerable noise, as long as the true trend was well represented. However, the interpolating function resulting from the ANN solution is a superposition of sigmoidal

functions and has the tendency to lead to artifacts in the derivatives, which manifest in slight, but undesirable bias in the smoothed traces.

Alternatives to ANNs are filters, such as the popular Kalman or Savitzky-Golay filter or the Whittaker filter which was proposed over eighty years ago [151]. Recently, Eilers presented a matrix form of this older implicit method, which was called a "perfect filter" [152]. Vilela and co-workers configured the Whittaker-Eilers smoother [153] by adapting Rényi's second-order entropy of the cross-validation error as the optimization criterion. The filter, implemented in the software AutoSmooth, can be used to extract signals and derivatives from time series with non-stationary noise structure [154].

4.3 Constraining the parameter search space

To ensure that the results of a parameter estimation fall within reasonable ranges it is often useful to constrain the optimization process, including guesses for the initial values, to suitable ranges or permitted extreme values for all parameters. For instance, in BST representations, the structural features of a system are mapped onto model parameters in a unique fashion, as described in Section 2.3.3. Therefore, if the network structure and regulation are known, one may be able to decide immediately whether the kinetic order of a variable X_j is positive, negative, or zero, depending on its influence (activation, inhibition, or no effect) on variable X_i. Furthermore, rate constants are always non-negative, and kinetic orders in BST pathway models are known to be real numbers with typical values between −1 and +2.

Some other supporting techniques aiming to reduce the parameter search space include the following. Kutalik *et al*. [83] characterized a one-dimensional basin of attraction containing the true optimum with minimal error. Tucker and Moulton [155] proposed a method based on interval analysis which allows exhaustive searches of the entire set of parameter values with a finite number of steps. Tucker *et al*. [156] used constraint propagation to find the possible ranges of parameter values, thus significantly constraining the parameter search space.

4.4 Reducing the complexity of the inference task

The typical approach of modeling is to collect network information and translate the wiring diagram into a symbolic model, which only contains a limited number of parameters since the biological systems are usually sparsely connected. However, when the topology of the system is unknown or only partially known, it appears that one must initiate the search with a full symbolic model with all parameters free. When the system is relatively small, it may be feasible to exhaust all possibilities and to find the optimum. However, when the number of variables and parameters grows, all methods of parameter estimation eventually run into problems of "combinatorial explosion," which makes the estimation process extremely difficult and the solutions problematic. This explosion can be tamed to some degree by constraining the connectivity within the system by systematically identifying the network structure or gradually "pruning" unlikely connection during optimization process. Of benefit in this context is the observation that biological systems are seldom fully connected and that indeed most nodes are only directly connected to a small number of other nodes [157,158]. These issues will be reviewed in detail in Section 5, in the context of structure identification techniques. In this section, we focus only on parameter pruning methods.

The rationale behind the pruning techniques is closely related to the characteristic of BST models. As briefly mentioned in Section 2.3.3, structure identification tasks can be translated into parameter estimation problems if the parameter values directly map to the network, as it is the case with BST representations. To recall this mapping, the kinetic orders f_{ij}, g_{ij} and h_{ij} for BST quantify the regulatory effect of variable X_j on a production or degradation term in the equation of variable X_i. If the magnitude of the corresponding kinetic orders are very small

or close to zero, the connection between variable X_j and the dynamics of X_i is likely to be negligible. Therefore, these low intensity connections can be purged during optimization, which not only helps to detect a reasonable and parsimonious model of the true pathway structure, but also reduces the parameter search space for further optimization.

The simplest manner of "pruning" a possibly highly connected network is to define a threshold for the absolute value of each type of parameter, below which values are set to zero [128, 159]. In addition, since the likelihood that a variable exists in both the production and degradation terms with non-zero values in the S-system model is low, the smaller of the kinetic orders is more likely to be zero and the value of the other one is adjusted accordingly [128].

Some authors have suggested more sophisticated methods for this pruning process. As an extension of the objective function for optimization, various articles have added to the residual error the sums of the absolute values of kinetic orders as a penalty term in the cost function. Thus, this basic pruning method for BST models penalizes all small kinetic orders, which have little effect on the system dynamics, and prevents the model from including false-positive interactions that unrealistically inflate the model [134,160]. To improve this condition further, Kimura and co-workers [82,124] introduced a penalty term that rearranged kinetic orders in ascending order based on their absolute values and eliminated those considered insignificant. Furthermore, accounting for the observation that very few factors modulate both the production and degradation of a specific variable, Noman and Iba [161] proposed an alternative representation of the penalty term.

No matter what kind of penalty term is chosen, pruning approaches pose an obvious challenge. Namely, the weighted coefficient in the penalty term needs to be carefully tuned since it affects the results of the structure identification task. So far there are no clear guidelines for setting suitable penalty weights. Stochastic ranking may be used to alleviate this difficulty since it aims to balance the error and penalty term in the objective function [162]. However, this method requires an additional parameter defining the probability of the error term for comparisons in ranking. Cho *et al.* [163] proposed a distinctly different way to retain the sparseness feature in biological pathways without adding extra terms to the objective function, which they coined S-tree representation. The S-tree is a tree representation of the S-system, where the number of sub-trees corresponds to the number of ordinary differential equations in the system. Each sub-tree is divided into two parts; the left part represents the production term and right part represents the degradation term. The depth of the S-tree is always three and the root node is at depth zero. Since S-tree modeling is intrinsically suitable for representing sparse networks, an S-tree together with genetic programming has the potential to infer network topology and find parameter values in a more efficient way without any *a priori* knowledge or adding a penalty term. To avoid assigning a coefficient weight to the penalty term, Liu and Wang recently proposed an alternative method based on multi-objective optimization [164,165]. Instead of minimizing the residual error using a single objective function either in concentrations or slopes, they minimized the concentration error, slope error, and interaction measure simultaneously. The authors proved that the algorithm guarantees the minimum solution for the constrained problem to achieve the minimum interaction network for the inference problem. The approach avoided assigning a penalty weight for sums of magnitude of kinetic orders.

Pruning methods are also used in optimization approaches for determining parameter values, as described in the next section.

4.5 Algorithms for determining optimal parameter estimates

The parameter estimation task is traditionally formulated as an optimization problem that minimizes an objective function measuring a generalized distance between experimental data and model predictions. The Euclidean distance is the most commonly used and often refers to

a least-squares error criterion. Other fitness evaluation methods include information based criteria [166,167]. Two objective functions are typically used for parameter estimation in metabolic models: a concentration error based objective function and a slope error based objective function (*e.g*. [140]). The concentration error based objective function is a straightforward calculation of the sum of squared distances between the metabolite measurements and the predictions. The simulation profiles are usually obtained by applying a numerical integration method to solve the differential equations like Eq. (1). The integration process can be computational costly, especially if the system is stiff (see Section 4.1). As an alternative, the slope error based objective function employs the decoupling technique as described in Section 4.1 and uses the slope information for evaluating fitness of the function. That is, it calculates the sum of squared errors between the measured slopes from the raw data (or upon smoothing) and the predicted slopes.

The most prominent methods for parameter estimation from time series data can be grouped into gradient-based methods, stochastic search algorithms, and others that do not belong to the first two groups. These optimization methods are reviewed in the following paragraphs and summarized in Figure 4.

4.5.1 Gradient-based algorithms—The most natural choice for estimating parameter values is presumably a gradient based regression, and many of the commercial methods of this type have been applied to metabolic models. Among these are Gauss-Newton and Levenberg-Marquardt methods, which are included in all major software packages of the field and will not be reviewed *per se* [168–170]. A comparison of some of these methods in the context of pathway models can be found in [50]. Marino and Voit [171] proposed a gradient based algorithm for finding the parameter values using BST models that comprises three modules in a novel fashion: model generation, parameter estimation for model fitting, and model selection. First, plausible initial models are generated in a step-by-step manner, upon decoupling and limiting connectivity (see Section 5.2 for detail). Secondly, each differential equation is fitted separately using the Levenberg-Marquardt method while replacing the other variables with raw data of smoothed traces. In the third phase, the model is compared to earlier, simpler models, and a statistical test decides whether the increased complexity of the model structure is warranted, as judged by the residual error. If the improvement is significant, a new, more complex model is generated, and this process is iterated until further advancements become insignificant.

Kutalik *et al*. [83] proposed an intriguing Newton-flow optimization method for parameter estimation in S-system models. The method starts with decoupling the differential equations and setting up an objective function for each equation. The next step is to select suitable start guesses and bounds for parameters and run a Newton method to obtain several points in the parameter space that correspond to reasonable, coarse solutions. The authors found that this space of coarse solutions contains a one-dimensional attractor. Standard regression allowed them to estimate the parameters of this attractor. Afterward, the Newton method was performed again using the initial guesses lying on the estimated attractor to find the true optimal of the parameter values. The interesting feature of this method is that most (or maybe even all) good parameter solutions seem to lie on one-dimensional manifolds within the high-dimensional parameter space. Optimization along this curve is comparatively easy. A potential problem of the method is that the original initial guesses for the parameters must lie within the basin of attraction of the one-dimensional manifold. Otherwise, each run may lead to disjoint sections of the parameter space.

Because biological systems are usually nonlinear, the problem of parameter estimation can be stated as a nonlinear programming problem (NLP) subject to nonlinear differential-algebraic constraints [172]. Because of its nonlinear and constrained nature, this inverse problem is

usually non-convex. Therefore, most of the traditional nonlinear algorithms involving gradient methods run the risk of getting trapped in local optima, depending upon the degree of system nonlinearity and the initial starting point [127].

4.5.2 Stochastic search algorithms—Several classes of stochastic methods are available for global optimization. They include evolutionary computation, simulated annealing, adaptive stochastic methods, clustering methods, and other meta-heuristics, such as ant colony optimization and particle swarm optimization. These algorithms have been applied to parameter estimation tasks with the goal of finding global solutions, especially in the context of identifying the structures of gene regulatory networks [172].

Evolutionary computation (EC) techniques, also known as biologically inspired methods, include genetic algorithms, evolutionary programming, evolution strategies, genetic programming, as well as their variants. They are attractive because they have a high potential of finding (at least the approximate locations of) global optima. Genetic algorithms (GA) have been shown to be very useful and practical in parameter estimations of biological systems (*e.g*. [150,160,172,173]). Using a conventional simple genetic algorithm (SGA), Tominaga *et al.* inferred parameter values of a small network, but only with a very limited number of parameters, and the convergence rate was low [174]. SGA typically has two problems: early convergence in the fast stage of the search and evolutionary stagnation in the last stage. Kikuchi *et al.* [134] enhanced SGA by using a more robust real coded genetic algorithm (RCGA) and improved the conventional cost function by adding a penalty term to prune unlikely connections in the investigated gene network using the S-system formalism. In addition, they employed a novel crossover method and introduced a gradual optimization strategy in the procedure. The results showed that the algorithm successfully inferred the network structure with faster convergence rate, optimization speed, and with more parameters predicted correctly, compared to the traditional GA. However, the approach turned out to be computationally very costly because of numerical integration of the entire system of differential equations (see Section 4.1).

Other modifications were made to improve the efficiency of SGA using time series data in S-system form. Examples include: a hybrid algorithm of SGA with a Modified Powell method [175]; a hybrid algorithm of SGA for static Boolean networks applied to an S-system with steady state and temporal data [81]; and a combination of RCGAs with unimodal normal distribution crossover and minimal generation gap to optimize parameters in S-systems [176–178]. Daisuke and Horton optimized an S-system model with a distributed genetic algorithm with “scale-free” properties [179]. Ho *et al.* [180] proposed an intelligent two-stage evolutionary algorithm (iTEA), which used an intelligent to solve decomposed ODEs independently, then combined all solutions from each subtask and used an orthogonal experimental design-based simulated annealing algorithm to refine the solution.

Spieth and coworkers [181,182] proposed a memetic algorithm (MA) consisting of two parts: a local search with an evolutionary strategy (ES) for parameter estimation, and a global GA based search framework for structure identification, where the former is embedded within the latter part. They tested the algorithm on an S-system model and the results showed that MA was better suited for inferring genetic networks than a standard ES or GA. In follow-up work, they showed that feedback from the local search to the GA based search can further improve the performance of MA [183].

Kimura *et al.* [124] used an evolutionary algorithm called Genetic Local Search with distance independent Diversity Control (GLSDC) and combined it with a decomposition strategy to estimate S-system models of gene regulatory networks. The proposed method included an estimation technique for the initial gene expression levels and enabled the reconstruction of medium-scale genetic networks with noisy data. They also showed that the combination with

a cooperative co-evolution algorithm can further improve the accuracy of prediction [82]. Okamoto's group also proposed evolutionary search techniques, such as the Network-Structure-Search Evolutionary Algorithm (NSS-EA) and a variant, the Grid-Oriented Genetic Algorithm Framework (GOGA Framework). They employed an S-system as the underlying mathematical model and used a GA as search engine to infer network structure [184–186].

Noman and co-workers recently incorporated their previously developed techniques into an improved memetic algorithm for inferring gene regulatory networks [161,166,187–189]. They used differential evolution (DE) along with a hill-climbing local-search method in their evolutionary algorithm. An information criterion-based fitness evaluation was introduced instead of the conventional least-squared errors approach.

Tsai and Wang [140] used hybrid differential evolution (HDE) for estimating a satisfactory, though not optimal solution, and then used the solution as the starting point for a gradient-based optimization method to obtain refined solutions. As described in Section 4.1, they used a modified collocation method to avoid direct numerical integration. In their recent work, they implemented HDE combined with a multiple-objective optimization approach (see Section 4.4 for review) to infer biochemical networks in S-system format [164].

Genetic programming (GP) has also been employed to discover the topology of metabolic pathway from time-series data (*e.g.* [190]). GP is an evolutionary algorithm that evolves mathematical expressions or computer programs. Traditionally, GP represents a mathematical expression or computer program as a tree structure, in which every tree node has an operator function and every terminal node has an operand. The general process of GP includes: initialization (randomly generate trees as individuals), evaluations (calculate fitness of each individual), selection (select individuals from the group base on probability), crossover (randomly select two individuals as parents and swap randomly chosen sub-trees of the parent trees), and mutation (such as insertion or deletion of terminal nodes) (*cf.* [191] for detail). The GP process makes mathematical expressions easy to evolve and evaluate. Therefore, in contrast to GA algorithms, which usually require defining equations before optimization, GP provides a general approach for finding arbitrary equations from time series data without specific knowledge of the equation. The ordinary GP is not always effective in finding the parameter values because the method relies mainly on the combination of randomly generated constants. Sagamoto and Iba [192] therefore used a least-mean-square (LMS) method along with ordinary GP to improve efficiency, using an S-system as one example. Their results showed that the fitness values improved faster in the early phase with the LMS method compared to the non-LMS method, since the former seemed to provide a better seed for the GP search.

Sugimoto and co-workers [193] implemented GP along with adding a penalty term to the cost function and introducing numeric mutations to the conventional procedure. They tested this method by predicting two equations of a metabolic reaction scheme regarding adenylate kinase and phosphofructokinase in Michaelis-Menten format, the equation of which is hard to derive if the underlying mechanism is not known. While their results showed that the algorithm can predict the equations with relatively simple forms, the method is already very time consuming for this relatively small system.

Kim *et al*. [194] adopted a symbolic pre-processing regression step in GP to avoid time consuming numerical integration, since the estimation of slopes for each data point in the time series can be obtained from the results of GP. Cho and co-workers [163] took advantage of the fact that GP has an evolving tree structure for given data and proposed S-tree based genetic programming for parameter estimation and structural identification in S-system models. As introduced in Section 4.4, this approach intrinsically accounts for the sparseness of the biological network. Therefore, even though no *a priori* knowledge about the network is known,

the S-tree based GP can still identify the underlying structure rather efficiently without adding a penalty term in the objective function.

The discussion in the previous paragraphs indicates that a considerable number of recent proposals applied evolutionary algorithms to tackle inverse tasks with BST models. So far no comprehensive comparison among these algorithms has characterized their relative efficiency, robustness, and accuracy. However, some more limited comparisons have been presented. Moles *et al*. [172] compared some stochastic global optimization methods using the case study of a biochemical model that consisted of 36 parameters and was formulated as a set of eight ODEs. This model was formulated in Michaelis-Menten representation, which could not take advantage of the highly structured format of BST representations. Spieth *et al*. [195] compared six evolutionary algorithms in three model frameworks: linear weight matrices, S-systems, and H-systems, where one fitness function was used to evaluate the convergence of algorithms. A comprehensive comparison of evolutionary algorithms is still needed.

Simulated annealing, colony optimization, and particle swarm optimization are also stochastic optimization methods. Simulated annealing (SA), a physically inspired method, was created to simulate the heating and cooling process of metal or glass, where atoms (solutions of the parameter estimation task) are allowed to leave their current state of low energy (fitness) during heating and have a chance of finding an even lower state (better fit) during cooling [196]. SA can behave as a global or local optimization search and automatically switches from a global to a local search when the “temperature” goes down. Gonzalez *et al*. [197] adapted SA for S-system parameter estimations from time series data. They tested the algorithm using three artificial datasets and assumed that the structure was either known or unknown and solved the entire set of ODEs per integration or upon decoupling. They also applied the algorithm to a real biological system.

Ant colony optimization (ACO) was inspired by the behavior of ants during the search for short paths between their colony and food sources, using pheromones that attract other ants, which then increase the amount of pheromone [198]. Over time, ever shorter paths (better solutions of the estimation task) become more popular. ACO is a probabilistic technique for solving computational problems that can be reduced to finding good paths through nodes in a graph. Zuñiga *et al*. [199] adapted ACO for S-system models by treating each metabolite as a node in a graph and inferring how other nodes were connected to it. They called their algorithm for identifying network structure a “discrete ACO.” As an extension of ACO they furthermore proposed a variant of an enhanced aggregation pheromone system (eAPS) for parameter estimation tasks involving S-systems, called a “continuous ACO.” The discrete ACO starts with a fully connected graph which corresponds to a set of equations where all variables are included in every equation. Their preliminary results showed that ACO produces good results when the test systems are very small, However, although the discrete ACO was able to eliminate some nodes (metabolites) from the graph in larger systems, it had problems eliminating unlikely connections and thus still produced an unreasonably large search space for the parameter estimation task. The authors concluded that the large search space might have been the reason for the continuous ACO to get trapped in local minima during the parameter estimation phase.

Particle swarm optimization (PSO) is a stochastic, population-based evolutionary computation algorithm. The original form of the PSO algorithm, which was motivated by social-psychological principles such as bird flocking and fish schooling, was first described by Eberhart and Kennedy [200]. In PSO, each potential solution is represented as a particle. A collection of potential solutions is called a swarm which consists of particles that fly around in a multidimensional search space. During flight, each particle adjusts its position according to its own experience and also collaborates with its neighboring particles through communication. When a particle encounters a promising solution, the area surrounding the

solution is further explored by the swarm. Therefore, PSO combines local search methods with global search methods. Naval *et al.* [201] adapted and refined PSO to scan the parameter space of a BST model.

4.5.3 Other algorithms—Some methods that aim to reduce the parameter search space using BST formalisms are described in Section 4.3 [83,155,156]. For linear parts of pathways, a technique of "peeling" terms [202] can be applied to models in BST to convert the nonlinear parameter estimation task into a series of linear regression tasks. Specifically, beginning with an equation that contains only one unknown power-law term, the differentials are substituted by slopes and the parameters of the unknown terms are estimated by linear regression. The results are fed into the equation of the subsequent metabolite, thereby making it amenable to linear regression as well, and this process is iterated to the end of the linear pathway.

As described in Section 4.5.1, the traditional gradient methods usually run the risk of getting trapped in local optima. To alleviate the problem, Polisetty *et al.* [203] proposed a branch-and-reduce algorithm to convert inverse problems involving GMA models into a convex optimization problem that is guaranteed to obtain the global solution within a predescribed parameter space. The major drawback of this method is its computational complexity.

Alternating regression (AR) [37] employs a decoupling technique for systems of differential equations and dissects the nonlinear parameter estimation task for S-systems into iterative steps of linear regression. The method is uniquely geared toward BST systems, because it utilizes the fact that power-law functions are linear in logarithmic space. AR is extremely fast in comparison to conventional methods and works well in many applications, if it converges. In cases where convergence is an issue, the fast speed renders it feasible to dedicate some computational effort to identifying suitable start values and search settings. AR is beneficial for the identification of system structure in S-systems as well. An extension of AR was successfully applied to S-distributions within the field of computational statistics [204].

As an extension of AR, eigenvector optimization (EO) [159] is based on a matrix formed from multiple regression equations of a decoupled S-system that is considered in logarithmic space. In contrast to AR, EO operates initially only on one term, whose parameter values are optimized completely before the complementary term is estimated. It was demonstrated that the EO algorithm converges fast and can be expected to converge in most cases, without necessarily requiring knowledge of the network structure. EO is easily extended to the optimization of network topologies with stoichiometric precursor-product constraints among equations.

Another recently proposed approach to metabolic systems estimation, called Dynamic Flux Estimation (DFE), consists of two distinct phases and applies particularly well to GMA models [21]. It corresponds to stoichiometric analysis, but does not consider systems in steady state but rather over a time horizon. The first phase attempts to establish a linear relationship among all fluxes in the system at each time point. Under certain rank conditions, these relationships form a matrix that can be solved uniquely. This first phase is entirely model-free and essentially assumption-free and includes a diagnosis of inconsistencies within the time series, and between the assumed system topology and the given data. The result of the first phase consists of numerical representations of all fluxes as they depend on the variables affecting them. The second, model-based phase addresses the mathematical formulation of these flux representations as explicit functions of the involved variables. Different from currently available methods, this phase allows quantitative diagnostics of whether the chosen mathematical representations are suitable. The two-phased approach thus permits rigorous, quantitative diagnoses of the data, the model structure, the assumptions made in the choice of flux representations, and of the various causes of residual errors. Preliminary results suggest that the DFE approach is more effective and robust than alternatives that are presently available,

if sufficient suitable data are available. Its combined model-free and model-based analyses reduce compensation of error between equations and between flux terms and promise significantly improved extrapolability toward new data or experimental conditions. Its diagnostic tools pinpoint causes of inadequate fits between model and data and suggest either changes in assumptions related to model choice or the use of data as un-modeled "off-line data." The main drawback of DFE is the requirement of rather comprehensive metabolic time series data, which however can be obtained in cases with already existing experimental methods. Furthermore, a direct application of DFE requires that the stoichiometric flux system is of full rank, which is usually not the case and requires additional "substitute information" [86]. Other issues needing refinement are related to missing data, missing flux information, error compensation among the parameters within a given flux, and ill-characterized systems topologies (see also Section 7).

Other methods which were developed recently for inverse problems in biological systems are discussed in [139,205–210]. However, these methods have not yet been implemented specifically for BST applications.

5 Inference of network structure

As mentioned in Section 1, the traditional approach of modeling begins with collecting network information that is translated into the design of a stoichiometric wiring diagram, which may then be converted into a fully kinetic metabolic pathway model, if desired. The translation and conversion more or less reflect the actual biological system as long as the input information is essentially correct and complete. In reality, information on network connectivity and regulation is often only partially known and seldom fully understood. As a consequence, the model design phase is subject to uncertainties, up to a point where it seems impossible to determine even the initial wiring diagram. Some of the identification methods discussed in this section ameliorate the situation and make it possible to deduce structure and regulation from time series data, at least in principle. Before we discuss these methods, it is beneficial to revisit aspects of the investigated system and the experimental data that have a direct effect on the complex identification task.

The need for valid system identification can be described in three aspects. First, wrong hypotheses regarding variables and interactions to be included in the model tend to lead to wrong interpretations of the results. Second, overly complex model representations may provide good fits to the observed time series data used for estimation but are unlikely to perform as well when tested against new datasets, due to over-fitting. Third, the inclusion of too many components and interactions in the model will eventually result in problems caused by combinatorial explosion, which means that any computational technique will ultimately be overwhelmed by the rapidly increasing number of equations, variables, and interactions between variables in large systems.

Fortunately, biology naturally offers a counteracting and very beneficial feature: namely the likelihood that a real biochemical, genetic, or proteomic network is fully connected is very low, because most metabolites (genes, proteins) are connected only to a limited number of other metabolites (genes, proteins); in fact, the vast majority of metabolites (genes, proteins) are involved in fewer than four or five processes each [157,211,212]. To take advantage of this fact of nature, it is therefore a desirable goal to precede any estimation attempt with a concerted effort to limit the number of candidate (structural and functional) connections within a system, as far as this is objectively possible. This *a priori* type of limitation can very significantly reduce the parameter space that must be searched, because structure identification and parameter estimation are closely related to each other, at least if canonical models are used.

In this section, we review some of the structure identification techniques; they can be categorized into two groups. First, model-free, coarse structure identification methods can be used to prescreen the particular situation at hand. These methods are based directly on the data and involve the determination of an estimated Jacobian matrix after small perturbations about the system's normal operating point, deductions from direct observation of the time profiles, a correlation-based approach, and a Bayesian network technique. The second group consists of model based structure identification methods, including "simple-to-general" and "general-to-specific" modeling strategies, as well as various additional methods using time series data within the BST framework. These methods and representative references are summarized in Figure 5.

5.1 Model-free structure identification approaches

5.1.1 Methods based on the Jacobian matrix—Much of the information necessary for identifying network structure depends on dynamic experiments. One type of such experiments is the measurement of transient responses of the system after small perturbations about the steady state. If the perturbations are small enough, the system can be expected to behave in a roughly linear fashion. Thus, the measurements may be used to populate the Jacobian matrix of the corresponding linearization, which then reveals the connectivity of the network. Over the past two decades, several proposals have been made to obtain the Jacobian matrix from experimental observations. Chevalier and co-workers [213] solved the Jacobian by applying multilinear least-square fitting to perturbed data. This approach is straightforward but very sensitive to noise and missing data points, because the crucially important differencing procedure is prone to generating large errors. To avoid instabilities due to numerical differentiation, the same group suggested using an integral representation, which expressed the solution in terms of eigenvectors and eigenvalues and solved the equation using nonlinear regression [213]. The advantage of this approach is that no differentiation is needed and hence the slopes need not be estimated. However, the drawback of this method is that the fit to a sum of exponentials with undetermined exponents is numerically somewhat problematic, and the nonlinear regression does not necessarily provide a solution which fits the data.

To overcome this difficulty, Sorribas *et al.* [214] suggested to reformulate the integral representation of the target function by reducing it to a multilinear regression problem. As the result, the eigenvalues of the Jacobian in the previous method can be easily calculated. However, the computation of eigenvalues is again rather sensitive to noise and rounding error, rendering the method not very reliable unless the multiplicities of the eigenvalues are exactly known. In order to avoid this problem, Díaz-Sierra and co-workers [215] proposed a variation to the previous methods, in which they directly obtained the Jacobian by expanding it in its Taylor-series without searching for eigenvalues. This methods yielded faster convergence.

All these methods are based on linear approximation, which is valid as long as the perturbation from the steady state remains relatively small. Thus, on one hand, the range of deviations needs to be small enough to yield a sufficiently accurate representation. On the other hand, the perturbation must be large enough to generate measurable responses. To alleviate this dilemma, Veflingstad *et al.* [216] suggested using the entire time course and fit the data in a piecewise linear fashion, using as an illustration example an S-system within BST. In the proposed method, the time series is subdivided into appropriate time intervals and the linearization is computed about a chosen operating point within each subset. Therefore, instead of focusing on one operating point, most reference states are different from the steady state. The results show that the piecewise approach is more likely to capture the relationship between variables in the system and can tolerate larger perturbations. The authors also showed that the collection of estimated coefficients resulting from different variations of linearization provided very strong clues about which variables were likely to be involved in a given equation and which

were not. These clues reflected likely parameter ranges or likely constraints on parameter values of the true model. A major drawback is that this method does not identify parameter values *per se*. For instance, as discussed in the original paper [216], it does not allow a distinction between various combinations of g_{ij} and h_{ij} in the S-system form because only their difference is being assessed as a single parameter in the linearization. However, this information is valuable. If additional information on the Jacobian matrix and both the concentration and fluxes at steady state are known, the difference between g_{ij} and h_{ij} can be directly calculated [217]. Also, if the difference has a magnitude that is significantly different from 0, it is likely that one of the kinetic orders is zero, because it is rare that a variable influences both production and degradation of the same variable. Therefore, if one can detect which connection may be omitted, the kinetic order can be computed straightforwardly.

Hatzimanikatis, Floudas and Bailey [218,219] indirectly contributed to the topic of structure identification per linearization by optimizing not only the production of yield in an S-system at steady state, as it has been done many times (*e.g.* [34,220,221]), but by also optimizing its regulatory structure. This numerical and structure optimization task led to a mixed integer linear programming (MILP) approach, for which standard software is available.

5.1.2 Direct observation—Unlike the previous methods for determining the Jacobian matrix by examining the linear properties on small amplitude perturbation near one or more operating points, the network connectivity can be deduced to some degree from direct observations on responses to perturbations of arbitrary amplitude made at different locations in the network. Vance and coworkers [222] proposed a strategy based on perturbing different components in a network and showed that relationships between the perturbed component and the remaining components may be deduced by observation of features in the response profile. These features include the order and size of the extreme values of the unperturbed components in response to the perturbed component, and the initial slopes of the time series at the perturbation. The former reflects the topological distances among the perturbed components and the remaining components in the network, while the latter reveals whether the components are directly affected by the perturbed variable or not. This distinction is accomplished by checking if the initial slopes are nonzero or zero upon perturbation. Vance and collaborators showed that this approach works well in some artificial networks including branching, feedback, and regulatory interactions. This method was also applied to an *in vitro* experiment with a glycolysis system, where the authors measured concentration changes in the reactor following impulse changes of different reaction metabolites [223]. From the experimental time series data the authors were able to identify some of the causal connectivities among the metabolites in the reaction pathway. Even though the method performed well in the synthetic time series and with experimental data from relatively small systems, this approach may not be applicable to more complicated networks, where the interpretation of profiles and the network reconstruction must be expected to be much harder. The emerging field of causality analysis [224,225] may be helpful for this type of analysis in the future.

5.1.3 Correlation-based approach—Some alternative approaches have been suggested for the reconstruction of chemical reaction networks. Arkin and co-workers [226,227] showed how correlations among components measured in the system may be used to infer or reconstruct a chemical reaction pathway. The approach, termed correlation metric construction (CMC), is based on the calculation and analysis of a time-lagged multivariate correlation function of time series data that are subjected to a series of random, large amplitude changes in the input concentration. The correlation information is used to construct the distance matrix and interpreted using a two-dimensional graph obtained with a projection technique called multidimensional scaling (MDS). The graph represents the connectivity and the strength of interactions among the species in the network. For instance, the shorter distances in the graph imply stronger connections while longer distances represent weaker interactions. The approach

was tested experimentally on a part of an *in vitro* glycolysis system containing eight enzymes and fourteen metabolites [227]. Along the same lines, Samoilov and collaborators [228] proposed two methods, entropy metric construction (EMC) and entropy reduction method (ERM), for the analysis of correlations between species from time series data and the inference of their underlying network.

5.1.4 Bayesian network approach—Another approach of network structure identification is statistical in nature and uses Bayesian ideas for assessing the probability that the dynamics of one metabolite directly depends on the dynamics of another metabolite. At the core of these methods is a Bayesian network [229], which is a graph model that represents a set of nodes (variables) and their conditional probabilistic dependencies upon each other. Its analysis permits the explicit detection of causal associations among nodes in the system, as long as there are no structural or regulatory cycles, for instance, in the form of material recycling or feedback signals. While the latter exclusions are clearly restrictive, Sachs and coworkers [66] successfully used Bayesian network methods to investigate the structure of a protein-signaling network from single cell flow cytometry data. The computational methods confirmed and elucidated most of the previously reported causalities and revealed new relationships between the involved signaling proteins. Bayesian network methods have not been applied extensively to metabolic pathway systems, but more often to genomic networks, where the task was to reconstruct networks of expression traits, or networks comprised of both expression and disease traits [67].

5.2 Model-based structure identification methods

The task of structure identification using model based approaches is very difficult for non-canonical models, because there are infinitely many nonlinear models that would have to be explored, unless some additional rationale could guide the model selection. Even within the comparatively limited area of metabolic modeling, the choices of combinations of rate functions would be daunting [25]. In any case, some types of "basis functions" are required for about any strategies of inferring network structures, as described in the following sections.

5.2.1 'Simple-to-general' and 'general-to-specific' modeling—As briefly mentioned in the introduction of this section, overly complex models may fit the data very well since increasing the complexity of the model naturally allows more freedom to provide a better fit to the data, for instance, in terms of the sum of squared errors. However, an over-inflated model typically does not perform well when tested on new data. This problem is known as over-fitting. One approach for restricting model complexity and to find the optimal model size is to add a penalty term to the cost function that is minimized. The optimal model can then be determined by finding the one that minimizes the aggregate cost function [230]. The consequent problem of using this approach is how to weigh data fit against model complexity. One approach, coined "simple to general," calls for starting with the simplest reasonable model and adding one term at a time until a minimal cost function is found (*e.g.* [231]). In the opposite direction, the "general to specific" strategy initially includes everything possible in the model and then gradually eliminates terms until the minimum in the cost function is found [232]. Crampin and co-workers [233,234] used these two approaches of model construction to extract kinetic information from time series data. Although their results suggested that the general-to-specific algorithm outperforms the simple-to-general approach, they indicated that when the number of chemical species included in the model is large (~10), the numbers of possible elementary reactions are massive, thus making the computation difficult if not infeasible. Therefore, it is desirable to limit the size of the basic set below a reasonable upper bound. This strategy appears to be reasonable, because genomic, metabolic, and proteomic networks are generally sparsely connected [157]. An idea similar to the simple-to-general strategy was discussed by Marino

and Voit [171], who addressed structure identification tasks beginning with the simplest types of S-system equations (see below).

5.2.2 Use of time series data—Model-based methods of structure identification are especially powerful if they are based on time series data. As discussed earlier, parameter estimation from time series data usually requires considerable computational effort, and this effort increases dramatically when the structure of the underlying system is unknown. The challenges of identifying the correct structure and regulation of a system strongly suggest using all preprocessing tools available to limit the analysis to the most likely connections in advance, reduce the search space and identify good initial guesses for the parameters.

For the identification of structure from time series data, BST models seems particularly useful, especially if not much specific information about the mechanistic processes within the biological network is available. The advantages and features of BST representations have been reviewed in Sections 2.3.3 and 4.4 and need no further description here.

In addition to the computational pruning techniques reviewed in Section 4.4, pruning can also be achieved based on biological insight. Almeida and Voit [147] suggested making maximal use of other *a priori* biological information that might be available in addition to the time series data. As a specific example, Voit and Savageau [136] analyzed a yeast fermentation system in several *a priori* possible variations that corresponded to hypotheses regarding the existence of specific processes and regulatory signals and studied the improvement in error with statistical methods.

In a more generic fashion of "inverse pruning," and pursuing the "specific to general" strategy, Marino *et al*. [171] proposed an algorithm based on reconstructing S-system equations in a gradual progression. Using the decoupling technique and focusing on one differential equation at a time, they began with equations with constant input and simple substrate driven degradation, obtained the best possible fit, and then gradually added other variables to the equation, always checking their statistical significance. Thus starting from the minimal (and most parsimonious) model, choosing a modest connectivity index, and increasing the number of variables step by step, until a maximally allowed level of connectivity was reached, they identified small pathway systems rather efficiently.

Daisuke and Horton [179] also utilized the "scale-free" property of networks [235,236] to restrict the connectivity in biological systems during optimization procedure. Their results showed that the restriction increased the conversion ratio while reducing the average number of generations and reducing both false positive and false negative estimations of links in the network. Zuñiga *et al*. [199] recently proposed applying ant colony optimization (ACO) to the network inference problem using the S-system formalism. Their preliminary results showed that, starting with a fully connected network, ACO was able to recover the connectivity of the network.

Kimura *et al*. [237] proposed a function approximation approach outside BST to infer reduced Normalized Gaussian network (NGnet) models of genetic networks from time-series data. Their results showed that the method successfully inferred the genetic network structure from artificial datasets with high specificity and sensitivity. The method was also tested on random genetic networks and actual biological data. The computational time of this method was shown to be much shorter than for other inference methods.

6 Toward a streamlined "work-flow" for inverse modeling

As described in Sections 4 and 5, many methods have been developed recently that attempt to solve parameter estimation and structure identification problems through inverse modeling

using the BST formalism. Most of these methods were developed to address the main problem of optimizing parameter values against observed time series data using gradient based methods, regression algorithms, or evolutionary approaches. Other methods were proposed as support algorithms, for instance, methods for avoiding the time consuming integration of differential equations, smoothing noisy data and estimating slopes, restricting the parameter search space, excluding unlikely connections within the network, or reducing the number of parameters to be estimated.

Many of the published papers used a combination of several methods to solve the inverse problem. For instance, they may have applied decoupling techniques along with various optimization algorithms, tried to reduce the number of parameters before estimating their values, or included several objective functions to constrain the solution space.

In spite of the considerable number of methods that have been proposed for inverse modeling using BST models, each method has its pros and cons and there always seem to be conditions and situations where one method works well and the other not so. In the end, there is currently no algorithm that is perfect, or even sufficiently effective, for the majority of realistic cases. Granted, each proposed algorithm showed effective solutions for particular cases and superiority in comparison to other methods. Nevertheless, most algorithms were only tested against synthetic time series data with respect to robustness and algorithmic efficacy, and it is known well that many such results fail when the same methods are applied to real experimental data. Furthermore, different authors used different benchmarking systems, so that it is hard to tell from the published results which algorithms are superior to the others, and under what conditions

The difficulty of obtaining fair comparisons is a result of the following five issues. First, as said before, different biological systems were used to demonstrate the usefulness of the algorithms. It is clear that different systems generate synthetic time series with distinctly different properties. For instance, they strongly affect the features of the data matrices that are the basis for subsequent computation. Depending on the specific data, the matrix may be ill-conditioned or exhibit collinearities between rows or columns, and this algebraic consequence has a direct effect on the efficacy, correctness and reliability of the tested algorithms. As a result, it is difficult to compare the algorithmic methods and simultaneously to discern how they are influenced by the features of the test system. Second, the numbers of time series points included for computation are different or unstated. Thus, the effects of data point inclusion or missing data on the algorithm are unclear, but they can affect the information criteria and the fitness score of the method. Third, the objective functions selected for the optimization vary and thereby prevent direct comparisons among algorithms. Fourth, the constraints on the parameter values are often different. This seemingly minor issue makes it difficult to tell if the algorithm converges since the boundaries are relatively close to the true optimum or because of the efficiency of the algorithm. Fifth, in addition to testing the methods using noise-free data, artificial “measurement errors” are introduced to examine if the algorithms can still find the correct parameter values. However, the way and extent of adding noise and the methods used for data smoothing often differ, which renders fair comparisons difficult.

A cursory comparison of parameter estimation algorithms in biochemical pathways has been published, but only two networks were considered and neither of them was implemented for BST applications [172].

6.1 Benchmarking framework

To address the problems indicated above, del Rosario and co-workers [238] recently proposed a project called MADMan (Munich-Atlanta-Diliman-Manila), which aims to compare the published parameter estimation algorithms using BST formalisms in a systematically way,

including the testing of the algorithms with the same variety of networks, uniform benchmarking bases, and standardized evaluation criteria. The goal of the benchmarking framework is to develop a strategy for choosing a set of candidate algorithms given a biochemical or gene regulatory network and experimental data. MADMan is an ongoing project that constitutes a huge task, which will require substantial effort and cooperation among the participating groups.

The direct comparison of various optimization algorithms and different data and settings will ultimately be the least biased strategy to determine which algorithms are better than the others. One must be aware though that speed (or lack) of convergence and unsatisfactory performance in terms of fitness are merely some of the issues that need to be analyzed for each optimization algorithm or computational software. Other features may contribute to the problem and its solutions as well, such as data related issues, model related issues, and mathematical issues, as reviewed in Section 3.4.

6.2 Work-flow strategy

While the MADMan project will attempt to clarify the applicability of methods under a wide range of conditions, we propose in this section a streamlined "work-flow" strategy for estimating parameter values in BST models with currently available methods. The work-flow diagram consists of a decision process based on potential problems that are encountered with some regularity. These include issues related to the time series data, model of choice, computational efficiency, and mathematical redundancy during the inverse modeling process. The work-flow also suggests relevant diagnostic tools or corresponding solutions. It addresses the main optimization algorithms as well as other supporting methods and diagnostic techniques, along with some assumptions and educated guesses that are required to estimate all parameter values of a system of realistic size.

6.2.1 Goals—The ultimate goal of inverse modeling is to find a mathematical model that describes the biological phenomenon and predicts situations that had not been used for model identification or data fitting with correctness, robustness, and efficiency. These standards may not always be fulfilled simultaneously, thereby requiring compromises. For instance, algorithms that find the optimal solution may be expensive in terms of computational time, whereas some of the fast algorithms may only be able to find coarse solutions.

The judgment on algorithms is comparatively easy when synthetic time series are used for testing, since the criterion of "correctness" is then automatically given and easy to assess by checking the fitness score, testing the validity of the inferred network structure, and comparing the estimates with the true model parameters. However, in reality, the "correct" model is not known and goodness of fit cannot always guarantee the reliability and applicability of the model. For instance, the model with the smallest residual error might *a priori* be deemed mathematically the "best" model. However, a small error does not necessarily imply that the model is the best choice for describing the biological system. In many actual cases, the "best model" tends to have over-fitting problems (see Section 5) and may not be able to extrapolate toward untested conditions when no extra constraints are introduced. Furthermore, a solution that fits the observed time series quite well may not necessarily be unique. Other solutions may exist, with distinctly different parameters, and with fits of a similar quality. In fact, instead of aiming to find "the one best" model, one might set the goal of every inverse modeling strategy as the task of determining *all* models that are consistent with the data within some acceptable error. The resulting candidate set of parameters may be clustered tightly or scattered throughout the search space. In either case, the diversity of possible solutions is helpful for exploring potential model structures and proposing possible causal relationships among the network components. Furthermore, the candidate models can be assessed with respect to stability,

sensitivity, gains, or other features that might shed light on the models and the investigated biological system [1,35]. Comparative simulations with the candidate models may identify one or the other model as more likely or suggest hypotheses or critical experiments that ultimately reveal to true composition of the system at hand.

6.2.2 Flow diagram of inverse modeling strategy—The proposed flow diagram for inverse modeling is shown in Figure 6. The global time series data are entered into a matrix, which is then screened and preprocessed with diagnostic and corrective tools (Step ①). For instance, if the variable traces have similarly shaped dynamics, they may be (approximately) collinear with each other. The calculation of the condition number or correlation coefficient can point to possible collinearities in the data matrix (Step ②). If the time traces are collinear, one may remove the model redundancy by pooling collinear variables or ignoring a subset of them, or merge constant variables with the rate constant (Step ③). If there is no collinearity, a symbolic mathematical model of the system can be derived based on the model of choice, without numerical specification of parameter values (Step ④). It has been shown that S-system and GMA representations in BST are good candidates for this propose. After setting up the full model, it is advisable to search for possible simplifications. For instance, if the network topology is known, a reduced symbolic model can be formulated with some parameters set to zero, in accordance with the network diagram (Step ⑤). If the system contains ubiquitous metabolites such as ATP, which are involved in dozens of reactions, partial modeling techniques may be applied. This technique retains these variables as input to the model, rather than explicitly modeling them (see Section 3.4.1 for details). The “off-line” nature of the ubiquitous variables further reduces the complexity of the symbolic model. Since fast optimization is desirable for the initial stage, even if it is coarse, it is beneficial to employ the decoupling technique, which converts the differential into algebraic equations. The decoupling step involves the measurement of slopes, which may be assessed directly or upon smoothing (Step ⑥). Once the symbolic model is decoupled, the parameters of each equation can be estimated by some fast optimization algorithm (Step ⑦). Alternating regression (AR) was shown to be one of the algorithms that work quite well for many S-system models. If AR converges, the resulting model is ready for further analysis and evaluation. If the initial guesses lead to lack of convergence, the algorithm is restarted with a different set of initial guesses. Another option for this step may be a collocation method. If the system topology is not known or only partially known, algorithms or techniques for inferring the network connectivity are applied. These include prior linearization of the system dynamics or sorting of parameter combinations by their empirical likelihood of inclusion in an equation (Step ⑧; see Section 5 for detail). If the network topology is not known, it is also necessary to choose an optimization algorithm that does not depend critically on topological information; a suitable algorithm for BST models is eigenvector optimization (EO) with prior decoupling (Steps ⑨ and ⑩). Algorithms permitting ill-characterized system topologies are usually combined with pruning methods that eliminate unlikely connections between network components and reduce the number of parameters to be estimated during the process of estimation. If the fast algorithms are not able to yield acceptable fittings, some other, computationally more expensive algorithms such as genetic algorithms or evolutionary approaches are applied (Step ⑪). If the outcome of the initial fitting is not acceptable, the optimized parameter values may be used as start values for subsequent refining algorithms. These are typically more costly and may lead to better solutions, although they are not necessarily always effective. A significant consequence and advantage of the combination of approaches is that the result often consists of multiple parameter sets that are all consistent with the data and that can lead to new, testable hypotheses and may offer guidance for further theoretical and experimental investigation (Step ⑫). It may be possible that the algorithms are not even able to produce acceptable fits (Step ⑬). We will discuss this situation in detail in Section 7. Once the initial models are obtained, the next step is dedicated to model analysis, including model diagnostic and cross validation

as described in Section 1. If the models are deemed reliable and appropriate for the purposes of the modeling effort, they can be used for applications and for gaining a deeper understanding of the biological phenomenon; specifically, they can be used to make predictions, generate new hypotheses, or guide the design of additional biological experiments (Step ⑭). In contrast, a model analysis that indicates lack of robustness or discrepancies between model and observations reveals potentially fundamental problems in the model. In this not so rare situation one needs to return to earlier steps of the modeling process and refine the model in an iterative manner. For instance, the modeler might need to discuss with the expert biologists how to obtain additional information or identify possible mistakes in the assumed model structure or missing reactions or signals in the pathways. It may also be useful to resample the data with jackknife or bootstrap methods and to redo the analysis in order to explore possible alternative solutions.

Most of the steps of the inverse modeling "work-flow" can be automated. For instance, it is relatively easy to check for collinearities between time series once the data matrix is ready (Step ②). The full symbolic model of the system can be derived directly and per computer if the number of variables in the model is known [1]; Step ④). The slopes of the time series can be estimated directly or upon smoothing, using various algorithms (Steps ⑥ and ⑨). The actual parameter optimization is possible with many algorithms based on the time series and the structure of the model (Steps ⑦, ⑩, and ⑪). The network topology can be inferred giving the data matrix and fully connected model (Step ⑧). These more or less automatic steps can be worked into a data pipeline, at least in principle. Other steps are not as straightforward and thus require work manual intervention. For instance, even though methods of matrix diagnostics in Step ② point to collinearities in the data matrix, pooling variables or reducing the redundancy in the model in Step ③ requires some thought regarding the location of the affected variables and their relationships in the pathway. Further model reduction includes the decision of which variables to model explicitly, in cases where the model contains highly connected metabolites. These "intelligent" steps are required to reduce the network topology and the corresponding symbolic model (Step ⑤). Thus, the entire process is not yet fully automated and may always require human supervision. Nonetheless, some tools like Best-Kit [239], Cadlive [240], BSTBox [1], and BioinformaticStation [241] are beginning to provide interfaces that facilitate some of the steps involved in metabolic modeling.

Once several candidate solutions are obtained, the immediate question is whether there are reasonable guidelines for choosing between them. Several scenarios are to be anticipated. If many well-fitting solutions, either found with different optimization methods or obtained using some re-sampling scheme, are clustered closely within the parameter space, the solutions are parametrically similar and the networks they represent are essentially the same or very similar in structure. In contrast, if the optimization yields distinctly different solutions, exhibiting essentially the same residual error, it is *a priori* difficult to decide which model is best. Recent results in one of our estimation studies showed that a single data set allowed multiple distinctly different numerical solutions, especially if constraints on kinetic orders were set loosely. This was not entirely surprising because even one-variable S-systems are flexible enough to permit different parameter sets generating very similar graphs (*cf.* [204,242–247]). Without additional information, each such parameter set is an equally valid solution since it fits the data essentially equally well. However, problems may arise if a "wrong" solution is used for extrapolation to new conditions, as we will discuss in Section 7. By basing the estimation on many data sets and experimentally testing the same pathway under different conditions, the problem can often be alleviated, because the use of several data sets clearly constrains the flexibility of the underlying model considerably.

In many cases, one will obtain several alternative solutions. In other cases, the opposite may be true: in spite of the many options outlined before, it is still possible that even a combination

strategy cannot find an acceptable fit. Potential reasons and suggested future work are discussed in Section 7.

7 Open issues

As mentioned in Section 3.4 and in the previous section, the challenges of inverse modeling can be classified into issues related to data, model structure, computation, and mathematical features of the representation. Most of the recent articles have acknowledged and discussed various computational issues in great detail and some have addressed data and model related issues. However, there has been little discussion of model validity and quality beyond residual errors, the conditions under which the models can be obtained, and diagnostic tools for non-convergence or for situations where models cannot even be obtained with any degree of reliability.

These open issues fall into two categories. First, even if the algorithms are able to find a set of candidate models, it is possible that none of these models is acceptable for one reason or another. For instance, it may happen that model diagnosis and simulation studies reveal that none of the models are stable, that they are all overly sensitive, or that they do not exhibit reliable predictive ability. Other problems are lacking model fit for data not used in the estimation and model failure in extrapolations. Second, it is possible that the algorithms are not even able to produce acceptable fits. In these cases, the failure is usually imputed to the computational algorithms themselves. However, the sources of the problem may lie in a combination of the alleged model structure, the particular data sets, and the computational methods, and it is advisable to extend the diagnosis beyond the algorithmic techniques.

Following are some of the issues that should be addressed to improve the validity and reliability of the estimated model, beyond residual errors and computational efficiency.

1. *Data related issues*: Even though good smoothing techniques can solve part of the problem of missing data points or time series, effective diagnostic tools for checking consistency within data are still needed. One special property in modeling metabolic networks is that the mass of metabolites is conserved during the reaction. Therefore, by accounting for material flows entering and leaving each metabolite pool, one may be able to identify flows which might have been unknown, considered unimportant, or difficult to measure in the experiment. Furthermore, methods for assessing whether residual errors are due to idiosyncrasies or noise in the data are needed. Also, statistical methods need to be developed for determining the necessary number and density of time points in each dynamic profile. This determination may require stronger, practically applicable definitions for the complexity of a dynamic response.
2. *Model related issues*: Traditionally, when a mathematical framework is chosen for modeling, the fluxes in the metabolic pathway are represented using the same basis functions, such as a Michaelis-Menten or power-law representation. However, it is possible that not all fluxes are appropriately modeled by the same format; an example is the substrate uptake step in bacteria, which is likely to follow different mechanisms than enzyme catalyzed reactions (*e.g.*, [21]). It is well known that all mathematical representations of biological processes are local approximations that are guaranteed only in the vicinity of operating point. If the metabolite concentrations do not fall within small ranges, the model may or may not properly represent the dynamics. This situation becomes particularly important when a single model is used for more than one set of time series, each of which represents different experimental conditions. Good criteria for determining the appropriateness of the chosen mathematical representations are still lacking.

Sections 4 and 5, as well as Table 3, indicated that many of the recent parameter estimation and structure identification methods were developed specifically for S-system models and that comparatively few studies targeted GMA models. As described in Section 2.3.3, the main reason is the highly structured format of S-system models, which sometimes allows easier and more specific methods of optimization (see, for instance [37,159,202]). Nonetheless, GMA models are considered to be closer to biochemical intuition than S-systems and are therefore often preferred by biochemists for modeling metabolic pathway systems. Thus, the development of inverse algorithm for GMA models would be highly desirable in the future.

3. *Mathematical issues*: It has been observed many times that several solutions may be found for a given dataset and a selected model. One reason may be mathematical redundancy. Redundancies may occur within or between fluxes and within or between equations. Compensation between fluxes can be identified to some degree if it is possible to estimate each flux in addition to estimating an entire equation [21]. Solutions for numerical compensation within a single flux are still needed. They seem to require data covering relatively wide ranges of variation, multiple datasets or additional information about some of the parameter values. In addition to numerical compensation it is possible that models contain Hamiltonians, conserved quantities or symmetry groups, which permit completely equivalent solutions with different parameter settings. Some of these issues have been discussed in the context of BST [129–131].

Finally, one should emphasize the need for obtaining reliable solutions within short periods of time. In some cases, only a single estimation of the system may be needed, and it may be acceptable if this estimation takes several hours. However, once the field moves toward "estimation on the fly," solutions must be obtained within a few minutes or, preferably, within seconds. The need for fast solutions becomes especially pertinent if biologists and modelers together engage in concept map modeling, which permits the conversion of hypothesized network diagrams into numerical mathematical models [1]. Specifically, based on the known or hypothesized connectivity and regulatory information regarding the investigated system, the biologist designs a concept map consisting of a connectivity diagram of processes comprising the system and including known or assumed regulatory features, and provides semi-quantitative information on stimuli and measured or expected responses of the system. The modeler converts this information through combined methods of forward and inverse estimation into a mathematical construct that can subsequently be used for typical model analyses and to generate and test new hypotheses. This conversion step, which includes parameter estimation from time series, needs to proceed fast in order to permit interactive work, in which the modeler runs simulations with the model and the biologist-modeler team collaboratively interprets the results and devises improved concept maps. Because this method heavily depends on the biologist's initial intuition and hypotheses, many iterations between hypothesis formulation and diagram-to-model conversion are needed, thus demanding fast solutions that might not be absolutely precise but allow the interactive exploration of complex biological systems.

Whether bottom-up, top-down or concept map modeling is the method of choice, we hope to have conveyed that the estimation of model parameters and the identification of structure and regulation of ill-characterized biological systems is a vibrant field that will continue to offer challenges to teams of biologists, mathematicians, computer scientists and modelers throughout the foreseeable future.

Acknowledgments

The authors are grateful to Dr. Siren Veflingstad and two anonymous reviewers for critically reading the manuscript and providing constructive suggestions. This work was supported in part by a National Heart, Lung and Blood Institute Proteomics Initiative (Contract N01-HV-28181; D. Knapp, PI), a Molecular and Cellular Biosciences Grant (MCB-0517135; E.O. Voit, PI) from the National Science Foundation, a grant from the National Institutes of Health (R01 GM063265; Y.A. Hannun, PI), and an endowment from the Georgia Research Alliance. The work was also in part funded by the BioEnergy Science Center (BESC), which is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring institutions.

References

1. Goel G, Chou IC, Voit EO. Biological systems modeling and analysis: A biomolecular technique of the twenty-first century. J Biomol Tech 2006;17:252–269. [PubMed: 17028166]
2. Voit, EO.; Schwacke, JH. Understanding through modeling. In: Konopka, AK., editor. Systems Biology: Principles, Methods, and Concepts. CRC Press/Taylor & Francis Books; Boca Raton: 2007. p. 27-82.
3. Veflingstad, SR.; Dam, P.; Xu, Y.; Voit, EO. Microbial pathway models. In: Xu, Y.; Gogarten, JP., editors. Computational Methods for Understanding Bacterial and Archaeal Genomes. Imperial College Press; London: 2008.
4. Wu J, Voit E. Hybrid modeling in biochemical systems theory by means of functional petri nets. J Bioinform Comput Biol 2009;7:107–134. [PubMed: 19226663]
5. Wu JL, Voit EO. Integrative biological systems modeling: challenges and opportunities. Frontiers of Computer Science in China. 2009in press
6. Vodovotz Y, Constantine G, Rubin J, Csete M, Voit EO, An G. Mechanistic simulations of inflammation: Current state and future prospects. Math Biosci 2009;217:1–10. [PubMed: 18835282]
7. Gavalas, GR. Nonlinear Differential Equations of Chemically Reacting Systems. Springer-Verlag; Berlin: 1968.
8. Heinrich, T.; Schuster, S. The Regulation of Cellular Systems. Chapman and Hall; New York: 1996.
9. Palsson, BØ. Systems Biology: Properties of Reconstructed Networks. Cambridge University Press; New York: 2006.
10. Stephanopoulos, G.; Aristidou, AA.; Nielsen, J. Metabolic Engineering: Principles and Methodologies. Academic Press; San Diego, CA: 1998.
11. Varma A, Palsson BØ. Metabolic flux balancing: basic concepts, scientific, and practical use. Bio/Technology 1994;12:994–998.
12. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Curr Opin Biotechnol 2003;14:491–496. [PubMed: 14580578]
13. Bono H, Ogata H, Goto S, Kanehisa M. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res 1998;8:203–210. [PubMed: 9521924]
14. Edwards JS, Palsson BØ. The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 2000;97:5528–5533. [PubMed: 10805808]
15. Forster J, Famili I, Fu P, Palsson BØ, Nielsen J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res 2003;13:244–253. [PubMed: 12566402]
16. Selkov E, Maltsev N, Olsen GJ, Overbeek R, Whitman WB. A reconstruction of the metabolism of Methanococcus jannaschii from sequence data. Gene 1997;197:GC11–26. [PubMed: 9332394]
17. Okamoto, M. Symposium on Cellular Systems Biology. National Chung Cheng University; Taiwan: 2008. System analysis of acetone-butanol-ethanol fermentation based on time-sliced metabolic flux analysis.
18. Teixeira AP, Santos SS, Carinhas N, Oliveira R, Alves PM. Combining metabolic flux analysis tools and 13C NMR to estimate intracellular fluxes of cultured astrocytes. Neurochem Int 2008;52:478–486. [PubMed: 17904693]
19. Yang C, Hua Q, Shimizu K. Quantitative analysis of intracellular metabolic fluxes using GC-MS and two-dimensional NMR spectroscopy. J Biosci Bioeng 2002;93:78–87. [PubMed: 16233169]

20. Vallino JJ, Stephanopoulos G. Metabolic flux distributions in Corynebacterium glutamicum during growth and lysine overproduction. Biotechnol Bioeng 1993;41:633–646. [PubMed: 18609599]
21. Goel G, Chou IC, Voit EO. System estimation from metabolic time series data. Bioinformatics 2008;24:2505–2511. [PubMed: 18772153]
22. Mahadevan R, Edwards JS, Doyle FJ. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. Biophysical Journal 2002;83:1331–1340. [PubMed: 12202358]
23. Covert MW, Palsson BØ. Constraints-based models: regulation of gene expression reduces the steady-state solution space. J Theor Biol 2003;221:309–325. [PubMed: 12642111]
24. Gombert AK, Nielsen J. Mathematical modelling of metabolism. Curr Opin Biotechnol 2000;11:180–186. [PubMed: 10753761]
25. Schulz, AR. Enzyme Kinetics: From Diastase to Multi-enzyme Systems. Cambridge University Press; Cambridge; New York: 1994.
26. Michaelis L, Menten ML. Die Kinetik der Invertinwirkung. Biochemische Zeitschrift 1913;49:333–369.
27. Savageau MA. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. J Theor Biol 1969;25:365–369. [PubMed: 5387046]
28. Savageau MA. The behavior of intact biochemical control systems. Curr Topics Cell Regulation 1972;6:63–129.
29. Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. Eur J Biochem 1974;42:89–95. [PubMed: 4830198]
30. Lineweaver H, Burk D. The determination of enzyme dissociation constants. J Amer Chem Soc 1934;56:658–666.
31. Voit EO, Sands PJ. Modeling forest growth I. Canonical approach. Ecological Modelling 1996;86:51–71.
32. Savageau MA. Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. J Theor Biol 1969;25:370–379. [PubMed: 5387047]
33. Savageau, MA. Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology. Addison-Wesley Pub. Co. Advanced Book Program; Reading, Mass: 1976.
34. Torres, NV.; Voit, EO. Pathway Analysis and Optimization in Metabolic Engineering. Cambridge University Press; Cambridge, U.K.: 2002.
35. Voit, EO. Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists. Cambridge University Press; Cambridge, UK: 2000.
36. Voit, EO., editor. S-System Approach to Understanding Complexity. Van Nostrand Reinhold; NY: 1991. Canonical Nonlinear Modeling.
37. Chou IC, Martens H, Voit EO. Parameter estimation in biochemical systems models with alternating regression. Theor Biol Med Model 2006;3:25. [PubMed: 16854227]
38. Savageau MA, Voit EO. Recasting nonlinear differential equations as S-systems: A canonical nonlinear form. Math Biosci 1987;87:83–115.
39. Voit EO. S-system modeling of complex systems with chaotic input. Environmetrics 1993;4:153–186.
40. Atkinson MR, Savageau MA, Myers JT, Ninfa AJ. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. Cell 2003;113:597–607. [PubMed: 12787501]
41. Savageau MA. Design principles for elementary gene circuits: Elements, methods, and examples. Chaos 2001;11:142–159. [PubMed: 12779449]
42. Vera J, Balsa-Canto E, Wellstead P, Banga JR, Wolkenhauer O. Power-law models of signal transduction pathways. Cellular Signalling 2007;19:1531–1541. [PubMed: 17399948]
43. Irvine DH, Savageau MA. Network regulation of the immune response: alternative control points for suppressor modulation of effector lymphocytes. J Immunol 1985;134:2100–2116. [PubMed: 2857745]
44. Irvine DH, Savageau MA. Network regulation of the immune response: modulation of suppressor lymphocytes by alternative signals including contrasuppression. J Immunol 1985;134:2117–2130. [PubMed: 2857746]

45. Schwacke JH, Voit EO. The potential for signal integration and processing in interacting MAP kinase cascades. J Theor Biol 2007;246:604–620. [PubMed: 17337011]
46. Hatzimanikatis V, Bailey JE. MCA has more to say. J Theor Biol 1996;182:233–242. [PubMed: 8944154]
47. Visser D, Heijnen JJ. The mathematics of metabolic control analysis revisited. Metab Eng 2002;4:114–123. [PubMed: 12009791]
48. Fell, DA. Understanding the Control of Metabolism. Portland Press; London: 1997.
49. Kacser H, Burns JA. The control of flux. Symp Soc Exp Biol 1973;27:65–104. [PubMed: 4148886]
50. del Rosario RC, Mendoza E, Voit EO. Challenges in lin-log modelling of glycolysis in *Lactococcus lactis*. IET Syst Biol 2008;2:136–149. [PubMed: 18537454]
51. Heijnen JJ. Approximative kinetic formats used in metabolic network modeling. Biotechnol Bioeng 2005;91:534–545. [PubMed: 16003779]
52. Wang FS, Ko CL, Voit EO. Kinetic modeling using S-systems and lin-log approaches. Biochem Eng J 2007;33:238–347.
53. Sorribas A, Hernandez-Bermejo B, Vilaprinyo E, Alves R. Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations. Biotechnol Bioeng 2007;97:1259–1277. [PubMed: 17187441]
54. Lotka, A. Elements of Physical Biology. Williams and Wilkins; Baltimore: 1925.
55. May, RE. Theoretical Ecology: Principles and Applications. Blackwell; Oxford: 1976.
56. Volterra V. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Mem R Accad dei Lincei 1926;2:31–113.
57. Hernandez-Bermejo B, Fairen V. Lotka-Volterra representation of general nonlinear systems. Math Biosci 1997;140:1–32. [PubMed: 9029910]
58. Peschel, M.; Mende, W. The Predator-Prey Model: Do we Live in a Volterra World?. Akademie-Verlag; Berlin: 1986.
59. Voit EO, Savageau MA. Equivalence between S-systems and Volterra-systems. Math Biosci 1986;78:47–55.
60. Savageau MA. Development of fractal kinetic theory for enzyme-catalysed reactions and implications for the design of biochemical pathways. Biosystems 1998;47:9–36. [PubMed: 9715749]
61. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 2002;9:67–103. [PubMed: 11911796]
62. D'Haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury. Pac Symp Biocomput 1999:41–52. [PubMed: 10380184]
63. Bower, JM.; Bolouri, H. Computational Modeling of Genetic and Biochemical Networks Computational Molecular Biology Series. The MIT Press; 2001.
64. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–113. [PubMed: 14735121]
65. Kauffman, SA. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press; New York: 1993.
66. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308:523–529. [PubMed: 15845847]
67. Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, Sachs JR, Schadt EE. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. PLoS Comput Biol 2007;3:e69. [PubMed: 17432931]
68. Savageau MA. Genetic regulatory mechanisms and the ecological niche of *Escherichia coli*. Proc Natl Acad Sci U S A 1974;71:2453–2455. [PubMed: 4601590]
69. Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. Nature 2000;403:335–338. [PubMed: 10659856]
70. Hasty J, Dolnik M, Rottschafer V, Collins JJ. Synthetic gene network for entraining and amplifying cellular oscillations. Phys Rev Lett 2002;88:148101. [PubMed: 11955179]
71. Hlavacek WS, Savageau MA. Subunit structure of regulator proteins influences the design of gene circuitry: analysis of perfectly coupled and completely uncoupled circuits. J Mol Biol 1995;248:739–755. [PubMed: 7752237]

72. Hlavacek WS, Savageau MA. Rules for coupled expression of regulator and effector genes in inducible circuits. J Mol Biol 1996;255:121–139. [PubMed: 8568860]
73. Hlavacek WS, Savageau MA. Completely uncoupled and perfectly coupled gene expression in repressible systems. J Mol Biol 1997;266:538–558. [PubMed: 9067609]
74. Savageau MA. A theory of alternative designs for biochemical control systems. Biomed Biochim Acta 1985;44:875–880. [PubMed: 4038287]
75. Neidhardt, FC.; Savageau, MA. Regulation beyond the operon. In: Neidhardt, FC.; Curtiss, R., III; Ingraham, JL.; Lin, ECC.; Low, KB.; Magasanik, B.; Reznikoff, W.; Riley, M.; Schaechter, M.; Umbarger, HE., editors. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press; Washington, D.C.: 1996. p. 1310-1324.
76. Savageau MA. Significance of autogenously regulated and constitutive synthesis of regulatory proteins in repressible biosynthetic systems. Nature 1975;258:208–214. [PubMed: 1105191]
77. Savageau MA. Design of molecular control mechanisms and the demand for gene expression. Proc Natl Acad Sci U S A 1977;74:5647–5651. [PubMed: 271992]
78. Savageau, MA. Models of gene function: general methods of kinetic analysis and specific ecological correlates. In: Blanch, HW.; Papoutsakis, ET.; Stephanopoulos, G., editors. Foundations of Biochemical Engineering: Kinetics and Thermodynamics in Biological Systems. American Chemical Society; Washington, D.C.: 1983. p. 3-25.
79. Savageau MA. Demand theory of gene regulation. I. Quantitative development of the theory. Genetics 1998;149:1665–1676. [PubMed: 9691027]
80. Savageau MA. Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of *Escherichia coli*. Genetics 1998;149:1677–1691. [PubMed: 9691028]
81. Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y. Development of a system for the inference of large scale genetic networks. Pac Symp Biocomput 2001:446–458. [PubMed: 11262963]
82. Kimura S, Ide K, Kashihara A, Kano M, Hatakeyama M, Masui R, Nakagawa N, Yokoyama S, Kuramitsu S, Konagaya A. Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. Bioinformatics 2005;21:1154–1163. [PubMed: 15514004]
83. Kutalik Z, Tucker W, Moulton V. S-system parameter estimation for noisy metabolic profiles using newton-flow analysis. IET Syst Biol 2007;1:174–180. [PubMed: 17591176]
84. Mao, F.; Wu, H.; Dam, P.; Chou, I-C.; Voit, EO.; Xu, Y. Prediction of biological pathways through data mining and information fusion. In: Xu, Y.; Gogarten, JP., editors. Computational Methods for Understanding Bacterial and Archaeal Genomes. Imperial College Press; London: 2008.
85. Voit EO. The dawn of a new era of metabolic systems analysis. Drug Discovery Today BioSilico 2004;2:182–189.
86. Voit EO, Goel G, Chou I-C, da Fonseca L. Estimation of Metabolic Pathway Systems from Different Data Sources. IET Systems Biol. accepted (2009)
87. Kanehisa, M. The KEGG database. Novartis Foundation Symposium; 2002. p. 91-101.
88. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32:D277–D280. [PubMed: 14681412]
89. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, Tissier C, Zhang P, Karp PD. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2006;34:D511–D516. [PubMed: 16381923]
90. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 2004;32:D431–D433. [PubMed: 14681450]
91. Shiraishi F, Savageau MA. The tricarboxylic-acid cycle in Dictyostelium discoideum. 1. Formulation of alternative kinetic representations. Journal of Biological Chemistry 1992;267:22912–22918. [PubMed: 1429641]
92. Torres NV. Modeling approach to control of carbohydrate-metabolism during citric-acid accumulation by Aspergillus-niger. 1. Model definition and stability of the steady-state. Biotechnol Bioeng 1994;44:104–111. [PubMed: 18618452]
93. Torres NV, Voit EO, Alcón CH. Optimization of nonlinear biotechnological processes with linear programming. Application to citric acid production in Aspergillus niger. Biotechnol Bioeng 1996;49:247–258. [PubMed: 18623575]

94. Cascante M, Curto R, Sorribas A. Comparative characterization of the fermentation pathway of *Saccharomyces cerevisiae* using biochemical systems theory and metabolic control analysis: steady-state analysis. Math Biosci 1995;130:51–69. [PubMed: 7579902]

95. Curto R, Sorribas A, Cascante M. Comparative characterization of the fermentation pathway of *Saccharomyces cerevisiae* using biochemical systems theory and metabolic control analysis: model definition and nomenclature. Math Biosci 1995;130:25–50. [PubMed: 7579901]

96. Sorribas A, Curto R, Cascante M. Comparative characterization of the fermentation pathway of *Saccharomyces cerevisiae* using biochemical systems theory and metabolic control analysis: model validation and dynamic behavior. Math Biosci 1995;130:71–84. [PubMed: 7579903]

97. Curto R, Voit EO, Cascante M. Analysis of abnormalities in purine metabolism leading to gout and to neurological dysfunctions in man. Biochem J 1998;329(Pt 3):477–487. [PubMed: 9445373]

98. Curto R, Voit EO, Sorribas A, Cascante M. Validation and steady-state analysis of a power-law model of purine metabolism in man. Biochem J 1997;324(Pt 3):761–775. [PubMed: 9210399]

99. Curto R, Voit EO, Sorribas A, Cascante M. Mathematical models of purine metabolism in man. Math Biosci 1998;151:1–49. [PubMed: 9664759]

100. Ferreira AE, Ponces Freire AM, Voit EO. A quantitative model of the generation of N(epsilon)-(carboxymethyl)lysine in the Maillard reaction between collagen and glucose. Biochem J 2003;376:109–121. [PubMed: 12911334]

101. Alves R, Herrero E, Sorribas A. Predictive reconstruction of the mitochondrial iron-sulfur cluster assembly metabolism: I. The role of the protein pair ferredoxin-ferredoxin reductase (Yah1-Axh1). Proteins: Structure Function and Bioinformatics 2004;56:354–366.

102. Alvarez-Vasquez F, Sims KJ, Cowart LA, Okamoto Y, Voit EO, Hannun YA. Simulation and validation of modelled sphingolipid metabolism in *Saccharomyces cerevisiae*. Nature 2005;433:425–430. [PubMed: 15674294]

103. Alvarez-Vasquez F, Sims KJ, Hannun YA, Voit EO. Integration of kinetic information on yeast sphingolipid metabolism in dynamical pathway models. J Theor Biol 2004;226:265–291. [PubMed: 14643642]

104. Alvarez-Vasquez F, Sims KJ, Voit EO, Hannun YA. Coordination of the dynamics of yeast sphingolipid metabolism during the diauxic shift. Theor Biol Med Model 2007;4:42. [PubMed: 17974024]

105. Klapa MI, Park SM, Sinskey AJ, Stephanopoulos G. Metabolite and isotopomer balancing in the analysis of metabolic cycles: I. Theory Biotechnol Bioeng 1999;62:375–391.

106. Ratcliffe RG, Shachar-Hill Y. Measuring multiple fluxes through plant metabolic networks. Plant J 2006;45:490–511. [PubMed: 16441345]

107. Wiechert W. 13C metabolic flux analysis. Metab Eng 2001;3:195–206. [PubMed: 11461141]

108. Wiechert W, Möllney M, Isermann N, Wurzel M, de Graaf AA. Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. Biotechnol Bioeng 1999;66:69–85. [PubMed: 10567066]

109. Alvarez-Vasquez F, Hannun YA, Voit EO. Dynamics of positional enrichment: Theoretical development and application to carbon labeling in *Zymomonas mobilis*. Biochem Eng J 2008;40:157–174. [PubMed: 19412323]

110. Voit EO, Alvarez-Vasquez F, Sims KJ. Analysis of dynamic labeling data. Math Biosci 2004;191:83–99. [PubMed: 15312745]

111. Kacser H, Burns JA. Molecular democracy: who shares the controls? Biochem Soc Trans 1979;7:1149–1160. [PubMed: 389705]

112. Sorribas A, Cascante M. Structure identifiability in metabolic pathways: parameter estimation in models based on the power-law formalism. Biochem J 1994;298(Pt 2):303–311. [PubMed: 8135735]

113. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol 2003;1:E5. [PubMed: 12929205]

114. Du X, Callister SJ, Manes NP, Adkins JN, Alexandridis RA, Zeng X, Roh JH, Smith WE, Donohue TJ, Kaplan S, Smith RD, Lipton MS. A computational strategy to analyze label-free temporal bottom-up proteomics data. J Proteome Res 2008;7:2595–2604. [PubMed: 18442284]

115. Neves AR, Ventura R, Mansour N, Shearman C, Gasson MJ, Maycock C, Ramos A, Santos H. Is the glycolytic flux in *Lactococcus lactis* primarily controlled by the redox charge? Kinetics of NAD (+) and NADH pools determined in vivo by 13C NMR. J Biol Chem 2002;277:28088–28098. [PubMed: 12011086]
116. Szyperski T. 13C-NMR, MS and metabolic flux balancing in biotechnology research. Q Rev Biophys 1998;31:41–106. [PubMed: 9717198]
117. Goodenowe, D. Metabolomic analysis with Fourier transform ion cyclotron resonance mass spectrometry. In: Goodacre, R.; Harrigan, GG., editors. Metabolite Profiling: Its Role in Biomarker Discovery and Gene Function Analysis. Kluwer Academic Publishers; Dordrecht, The Netherlands: 2003. p. 125-139.
118. Plumb RS, Stumpf CL, Gorenstein MV, Castro-Perez JM, Dear GJ, Anthony M, Sweatman BC, Connor SC, Haselden JN. Metabonomics: the use of electrospray mass spectrometry coupled to reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development. Rapid Commun Mass Spectrom 2002;16:1991–1996. [PubMed: 12362392]
119. Ostergaard S, Olsson L, Nielsen J. *In vivo* dynamics of galactose metabolism in *Saccharomyces cerevisiae*: metabolic fluxes and metabolite levels. Biotechnol Bioeng 2001;73:412–425. [PubMed: 11320512]
120. Theobald U, Mailinger W, Baltes M, Rizzi M, Reuss M. *In vivo* analysis of metabolic dynamics in *Saccharomyces cerevisiae*: I. Experimental observations. Biotechnol Bioeng 1997;55:305–316. [PubMed: 18636489]
121. Voit EO, Almeida J, Marino S, Lall R, Goel G, Neves AR, Santos H. Regulation of glycolysis in *Lactococcus lactis*: an unfinished systems biological case study. IEE Proceedings Systems Biology 2006;153:286–298. [PubMed: 16986630]
122. Voit EO, Marino S, Lall R. Challenges for the identification of biological systems from *in vivo* time series data. In Silico Biol 2005;5:83–92. [PubMed: 15972008]
123. Maki Y, Ueda T, Okamoto M, Uematsu N, Inamura Y, Eguchi Y. Inference of genetic network using the expression profile time course data of mouse P19 cells. Genome Inform 2002;13:382–383.
124. Kimura S, Hatakeyama M, Konagaya A. Inference of S-system models of genetic networks from noisy time-series data. Chem-Bio Informatics Journal 2004;4:1–14.
125. Savageau, MA. Enzyme kinetics *in vitro* and *in vivo*: Michaelis-Menten revisited. In: Bittar, EE., editor. Principles of Medical Biology. JAI Press Inc; Greenwich, Conneticut: 1995.
126. Hill CM, Waight RD, Bardsley WG. Does any enzyme follow the Michaelis-Menten equation? Mol Cell Biochem 1977;15:173–178. [PubMed: 887080]
127. Mendes P, Kell D. Non-linear optimization of biochemical pathways: Applications to metabolic engineering and parameter estimation. Bioinformatics 1998;14:869–883. [PubMed: 9927716]
128. Voit EO, Almeida J. Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics 2004;20:1670–1681. [PubMed: 14988125]
129. Voit EO. Symmetries of S-systems. Math Biosci 1992;109:19–37. [PubMed: 1591448]
130. Sands PJ, Voit EO. Flux-based estimation of parameters in S-systems. Ecol Modeling 1996;93:75–88.
131. Voit, EO. S-System Approach to Understanding Complexity. Van Nostrand Reinhold; NY: 1991. Algebraic properties of canonical forms, Canonical Nonlinear Modeling; p. 278-303.
132. Berg PH, Voit EO, White RL. A pharmacodynamic model for the action of the antibiotic imipenem on Pseudomonas aeruginosa populations *in vitro*. Bull Math Biol 1996;58:923–938. [PubMed: 8837524]
133. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol 2007;3:1871–1878. [PubMed: 17922568]
134. Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M. Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics 2003;19:643–650. [PubMed: 12651723]
135. Irvine DH, Savageau MA. Efficient solution of nonlinear ordinary differential equations expressed in S-system canonical form. SIAM J Numer Anal 1990;27:704–735.
136. Voit EO, Savageau MA. Power-law approach to modeling biological systems; II. Application to ethanol production. J Ferment Technol 1982;60:229–232.

137. Voit EO, Savageau MA. Power-law approach to modeling biological systems; III. Methods of analysis. J Ferment Technol 1982;60:233–241.

138. Matsubara Y, Kikuchi S, Sugimoto M, Tomita M. Parameter estimation for stiff equations of biosystems using radial basis function networks. BMC Bioinformatics 2006;7:230. [PubMed: 16643665]

139. Rank E. Application of Bayesian trained RBF networks to nonlinear time-series modeling. Signal Process 2003;83:1393–1410.

140. Tsai KY, Wang FS. Evolutionary optimization with data collocation for reverse engineering of biological networks. Bioinformatics 2005;21:1180–1188. [PubMed: 15513993]

141. de Boor, C. A Practical Guide to Splines. Springer-Verlag; New York: 1978.

142. de Boor, C.; Höllig, K.; Riemenschneider, SD. Box Splines. Springer-Verlag; New York; Hong Kong: 1993.

143. Green, PJ.; Silverman, BW. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman & Hall; London; New York: 1994.

144. Seatzu C. A fitting based method for parameter estimation in S-Systems. Dynam Systems Appl 2000;9:77–98.

145. Burden, RL.; Faires, JD. Numerical Analysis. PWS Publishing Co; Boston, MA: 1993.

146. Almeida JS. Predictive non-linear modeling of complex data by artificial neural networks. Curr Opin Biotechnol 2002;13:72–76. [PubMed: 11849962]

147. Almeida JS, Voit EO. Neural-network-based parameter estimation in S-system models of biological networks. Genome Inform 2003;14:114–123. [PubMed: 15706526]

148. Funahashi KI. On the approximate realization of continuous mappings by neural networks. Neural Networks 1989;2:183–192.

149. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989;2:359–366.

150. Mendes P, Kell DB. On the analysis of the inverse problem of metabolic pathways using artificial neural networks. Biosystems 1996;38:15–28. [PubMed: 8833745]

151. Whittaker ET. On a new method of graduation. Proceedings of the Edinburgh Mathematical Society 1923;41:63–75.

152. Eilers PHC. A perfect smoother. Analytical Chemistry 2003;75:3631–3636. [PubMed: 14570219]

153. Vilela M, Borges CC, Vinga S, Vasconcelos AT, Santos H, Voit EO, Almeida JS. Automated smoother for the numerical decoupling of dynamics models. BMC Bioinformatics 2007;8:305. [PubMed: 17711581]

154. Vilela, M.; Borges, CC.; Vinga, S.; Vasconcelos, AT.; Santos, H.; Voit, EO.; Almeida, JS. Automated smoother for the numerical decoupling of dynamics models. http://autosmooth.sourceforge.net/(2007)

155. Tucker W, Moulton V. Parameter reconstruction for biochemical networks using interval analysis. Reliable Computing 2006;12:1–14.

156. Tucker W, Kutalik Z, Moulton V. Estimating parameters for generalized mass action models using constraint propagation. Math Biosci 2007;208:607–620. [PubMed: 17306307]

157. Jeong H, Tombor B, Albert R, Oltval ZN, Barabási AL. The large-scale organization of metabolic networks. Nature 2000;407:651–654. [PubMed: 11034217]

158. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. Bioessays 1998;20:433–440. [PubMed: 9670816]

159. Vilela M, Chou IC, Vinga S, Vasconcelos AT, Voit EO, Almeida JS. Parameter optimization in S-system models. BMC Syst Biol 2008;2:35. [PubMed: 18416837]

160. Voit, EO.; Almeida, JS. Dynamic profiling and canonical modeling: Powerful partners in metabolic pathway identification. In: Goodacre, R.; Harrigan, GG., editors. Metabolite Profiling: Its Role in Biomarker Discovery and Gene Function Analysis. Kluwer Academic Publishing; Dordrecht, The Netherlands: 2003.

161. Noman N, Iba H. Reverse engineering genetic networks using evolutionary computation. Genome Inform 2005;16:205–214. [PubMed: 16901103]

162. Runarsson TP, Yao X. Stochastic ranking for constrained evolutionary optimization. IEEE Transactions on Evolutionary Computation 2000;4:284–294.
163. Cho DY, Cho KH, Zhang BT. Identification of biochemical networks by S-tree based genetic programming. Bioinformatics 2006;22:1631–1640. [PubMed: 16585066]
164. Liu PK, Wang FS. Inference of biochemical network models in S-system using multiobjective optimization approach. Bioinformatics 2008;24:1085–1092. [PubMed: 18321886]
165. Liu PK, Wang FS. Inverse problems of biological systems using multi-objective optimization. Journal of the Chinese Institute of Chemical Engineers 2008;39:399–406.
166. Noman, N.; Iba, H. Inference of genetic networks using S-system: information criteria for model selection; Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation (GECCO'06); Seattle, Washington, USA: ACM Press; 2006. p. 263-270.
167. Shin A, Iba H. Construction of genetic network using evolutionary algorithm and combined fitness function. Genome Inform 2003;14:94–103. [PubMed: 15706524]
168. Björck, A. Numerical Methods for Least Squares Problems. SIAM; Philadelphia, PA: 1996.
169. Fletcher, R. Practical Methods of Optimization. Wiley; New York: 1987.
170. Nocedal, J.; Wright, SJ. Numerical Optimization. Springer; New York: 1999.
171. Marino S, Voit EO. An automated procedure for the extraction of metabolic network information from time series data. J Bioinform Comput Biol 2006;4:665–691. [PubMed: 16960969]
172. Moles CG, Mendes P, Banga JR. Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res 2003;13:2467–2474. [PubMed: 14559783]
173. Park LJ, Park CH, Park C, Lee T. Application of genetic algorithms to parameter estimation of bioprocesses. Med Biol Eng Comput 1997;35:47–49. [PubMed: 9136190]
174. Tominaga, D.; Koga, N.; Okamoto, M. Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. Proceedings of the Genetic and Evolutionary Computation Conference; 2000. p. 251-258.
175. Okamoto M, Nonaka T, Ochiai S, Tominaga D. Nonlinear numerical optimization with use of a hybrid Genetic Algorithm incorporating the Modified Powell method. Applied Mathematics and Computation 1998;91:63–72.
176. Nakatsui M, Ueda T, Okamoto M. Integrated system for inference of gene expression network. Genome Inform 2003;14:282–283.
177. Ueda T, Koga N, Okamoto M. Efficient numerical optimization technique based on real-coded genetic algorithm. Genome Inform 2001;12:451–453.
178. Ueda T, Ono I, Okamoto M. Development of system identification technique based on real-coded genetic algorithm. Genome Inform 2002;13:386–387.
179. Daisuke T, Horton P. Inference of scale-free networks from gene expression time series. J Bioinform Comput Biol 2006;4:503–514. [PubMed: 16819798]
180. Ho SY, Hsieh CH, Yu FC, Huang HL. An intelligent two-stage evolutionary algorithm for dynamic pathway identification from gene expression profiles. IEEE/ACM Trans Comput Biol Bioinform 2007;4:648–660. [PubMed: 17975275]
181. Spieth, C.; Streichert, F.; Speer, N.; Zell, A. A memetic inference method for gene regulatory networks based on S-Systems. Congress on Evolutionary Computation 2004 (CEC2004); 2004. p. 152-157.
182. Spieth, C.; Streichert, F.; Speer, N.; Zell, A. Genetic and Evolutionary Computation-GECCO 2004 (LNCS). Springer; Berlin/Heidelberg: 2004. Optimizing topology and parameters of gene regulatory network models from time-series experiments; p. 461-470.
183. Spieth, C.; Streichert, F.; Supper, J.; Speer, N.; Zell, A. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE Press; 2005. Feedback memetic algorithms for modeling gene regulatory networks; p. 61-67.
184. Imade, H.; Mizuguchi, N.; Ono, I.; Ono, N.; Okamoto, M. In: Konagaya, A.; Satou, K., editors. "Gridifying" an evolutionary algorithm for inference of genetic networks using the improved GOGA framework and its performance evaluation on OBI grid; Grid Computing in Life Science: First International Workshop on Life Science Grid, LSGRID 2004; Kanazawa, Japan. May 31–June 1, 2004; Berlin/Heidelberg: Springer; 2005. p. 171-186.

185. Morishita, R.; Imade, H.; Ono, I.; Ono, N.; Okamoto, M. Finding multiple solutions based on an evolutionary algorithm for inference of genetic networks by S-system. Congress on Evolutionary Computation 2003 (CEC2003); 2003. p. 615-622.

186. Ono, I.; Seike, Y.; Morishita, R.; Ono, N.; Nakatsui, M.; Okamoto, M. An evolutionary algorithm taking account of mutual interactions among substances for inference of genetic networks. Congress on Evolutionary Computation 2004 (CEC2004); 2004. p. 2060-2067.

187. Noman, N.; Iba, H. Inference of gene regulatory networks using s-system and differential evolution. Proceedings of the 2005 Conference on Genetic and Evolutionary Computation; Washington D.C.. 2005. p. 439-446.

188. Noman, N.; Iba, H. Enhancing differential evolution performance with local search for high dimensional function optimization. Proceedings of the 2005 Conference on Genetic and Evolutionary Computation; Washington D.C.. 2005. p. 967-974.

189. Noman N, Iba H. Inferring gene regulatory networks using differential evolution with local search heuristics. IEEE/ACM Trans Comput Biol Bioinform 2007;4:634–647. [PubMed: 17975274]

190. Koza JR, Mydlowec W, Lanza G, Yu J, Keane MA. Reverse engineering of metabolic pathways from observed data using genetic programming. Pac Symp Biocomput 2001:434–445. [PubMed: 11262962]

191. Koza, JR. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press; Cambridge, MA: 1992.

192. Sakamoto, E.; Iba, H. Inferring a system of differential equations for a gene regulatorynetwork by using genetic programming. Proceedings of the 2001 Congress on Evolutionary Computation (CEC2001); IEEE Press, Seoul, South Korea. 2001. p. 720-726.

193. Sugimoto M, Kikuchi S, Tomita M. Reverse engineering of biochemical equations from time-course data by means of genetic programming. Biosystems 2005;80:155–164. [PubMed: 15823414]

194. Kim, K-Y.; Cho, D-Y.; Zhang, B-T. Multi-stage evolutionary algorithms for efficient identification of gene regulatory networks. EvoWorkshops 2006; Springer; 2006. p. 45-56.

195. Spieth, C.; Worzischek, R.; Streichert, F. Comparing evolutionary algorithms on the problem of network inference. Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation (GECCO'06); ACM Press, Seattle, Washington, USA. 2006. p. 305-306.

196. Salamon, P.; Sibani, P.; Frost, R. Facts, Conjectures, and Improvements for Simulated Annealing. SIAM; New York: 2002.

197. Gonzalez OR, Kuper C, Jung K, Naval PC Jr, Mendoza E. Parameter estimation using Simulated Annealing for S-system models of biochemical networks. Bioinformatics 2007;23:480–486. [PubMed: 17038344]

198. Dorigo, M.; Di Caro, G. Ant colony optimization: a new meta-heuristic. Proceedings of the 1999 Congress on Evolutionary Computation (CEC1999); Washington, D.C.. 1999. p. 1470-1477.

199. Zuñiga, PC.; Pasia, J.; Adorna, H.; del Rosario, RCH.; Naval, P. An ant colony optimization algorithm for parameter estimation and network inference problems in S-system models. International Conference on Molecular Systems Biology 2008 (ICMSB08); Manila, Philippines. 2008. p. 105-106.

200. Eberhart, R.; Kennedy, J. A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micro Machine and Human Science (MHS'95); 1995. p. 39-43.

201. Naval, PC.; Sison, LG.; Mendoza, E. Metabolic network parameter inference using particle swarm optimization. International Conference on Molecular Systems Biology 2006 (ICMSB06); Munich, Germany. 2006.

202. Lall R, Voit EO. Parameter estimation in modulated, unbranched reaction chains within biochemical systems. Comput Biol Chem 2005;29:309–318. [PubMed: 16213792]

203. Polisetty PK, Voit EO, Gatzke EP. Identification of metabolic system parameters using global optimization methods. Theor Biol Med Model 2006;3:4. [PubMed: 16441881]

204. Chou IC, Martens H, Voit EO. Parameter estimation of S-distributions with alternating regression. Stat Operations Res Transactions (SORT) 2007;31:55–74.

205. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci U S A 2003;100:15522–15527. [PubMed: 14673099]

206. Srividhya J, Crampin EJ, McSharry PE, Schnell S. Reconstructing biochemical pathways from time course data. Proteomics 2007;7:828–838. [PubMed: 17370261]

207. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. Nature 2002;420:190–193. [PubMed: 12432396]

208. Thomas R, Mehrotra S, Papoutsakis ET, Hatzimanikatis V. A model-based optimization framework for the inference on gene regulatory networks from DNA array data. Bioinformatics 2004;20:3221–3235. [PubMed: 15247105]

209. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC. gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. Metab Eng 2005;7:128–141. [PubMed: 15781421]

210. Yeung MK, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci U S A 2002;99:6163–6168. [PubMed: 11983907]

211. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. Science 2002;298:824–827. [PubMed: 12399590]

212. Wagner A, Fell DA. The small world inside large metabolic networks. Proc Biol Sci 2001;268:1803–1810. [PubMed: 11522199]

213. Chevalier T, Schreiber I, Ross J. Toward a Systematic Determination of Complex Reaction Mechanisms. J Phys Chem 1993;97:6776–6787.

214. Sorribas A, Lozano JB, Fairén V. Deriving chemical and biochemical model networks from experimental measurements. Recent Res Devel Phys Chem 1998;2:553–573.

215. Díaz-Sierra R, Lozano JB, Fairén V. Deduction of chemical mechanisms from the linear response around steady state. J Phys Chem 1999;103:337–343.

216. Veflingstad SR, Almeida J, Voit EO. Priming nonlinear searches for pathway identification. Theor Biol Med Model 2004;1:8. [PubMed: 15367330]

217. Kitayama T, Kinoshita A, Sugimoto M, Nakayama Y, Tomita M. A simplified method for power-law modelling of metabolic pathways from time-course data and steady-state flux profiles. Theor Biol Med Model 2006;3:24. [PubMed: 16846504]

218. Hatzimanikatis V, Floudas CA, Bailey JE. Optimization of regulatory architectures in metabolic reaction networks. Biotechnol Bioeng 1996;52:485–500. [PubMed: 18629921]

219. Hatzimanikatis V, Floudas CA, Bailey JE. Analysis and design of metabolic reaction networks via mixed-integer linear optimization. AIChE Journal 1996;42:1277–1292.

220. Regan L, Bogle IDL, Dunhill P. Simulation and optimization of metabolic pathways. Computers Chem Engng 1993;17:627–637.

221. Voit EO. Optimization in integrated biochemical systems. Biotechnol Bioeng 1992;40:572–582. [PubMed: 18601153]

222. Vance W, Arkin A, Ross J. Determination of causal connectivities of species in reaction networks. Proc Natl Acad Sci U S A 2002;99:5816–5821. [PubMed: 11983885]

223. Torralba AS, Yu K, Shen P, Oefner PJ, Ross J. Experimental test of a method for determining causal connectivities of species in reactions. Proc Natl Acad Sci U S A 2003;100:1494–1498. [PubMed: 12576555]

224. Pearl, J. Causality: Models, Reasoning, and Inference. Cambridge University Press; 2000.

225. Spirtes, P.; Glymour, C.; Scheines, R. Causation, Prediction, and Search. Springer-Verlag; New York: 1993.

226. Arkin A, Ross J. Statistical construction of chemical-reaction mechanisms from measured time-series. J Phys Chem 1995;99:970–979.

227. Arkin A, Shen PD, Ross J. A test case of correlation metric construction of a reaction pathway from measurements. Science 1997;277:1275–1279.

228. Samoilov M, Arkin A, Ross J. On the deduction of chemical reaction pathways from measurements of time series of concentrations. Chaos 2001;11:108–114. [PubMed: 12779446]

229. Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann; San Mateo, CA: 1988.

230. Akaike H. New look at statistical-model identification. IEEE Transactions on Automatic Control 1974;AC19:716–723.

231. Judd K, Mees A. On selecting models for nonlinear time-series. Physica D 1995;82:426–444.

232. Hendry, DF.; Krolzig, HM. New developments in automatic general-to-specific modelling. In: Stigum, BP., editor. Econometrics and the Philosophy of Economics. Princeton University Press; Princeton: 2003.

233. Crampin, EJ.; McSharry, PE.; Schnell, S. Lecture Notes in Artificial Intelligence. Springer-Verlag; 2004. Extracting biochemical reaction kinetics from time series data; p. 329-336.

234. Crampin EJ, Schnell S, McSharry PE. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. Prog Biophys Mol Biol 2004;86:77–112. [PubMed: 15261526]

235. Barabási AL, Albert R, Jeong H, Bianconi G. Power-law distribution of the World Wide Web. Science 2000;287:2115.

236. Podani J, Oltvai ZN, Jeong H, Tombor B, Barabasi AL, Szathmary E. Comparable system-level organization of Archaea and Eukaryotes. Nat Genet 2001;29:54–56. [PubMed: 11528391]

237. Kimura S, Sonoda K, Yamane S, Maeda H, Matsumura K, Hatakeyama M. Function approximation approach to the inference of reduced NGnet models of genetic networks. BMC Bioinformatics 2008;9:23. [PubMed: 18194576]

238. del, Rosario RCH.; Echavez, MT.; de Paz, MT.; Zuñiga, PC.; Bargo, MCR.; Talaue, CO.; Arellano, C.; Pasia, JM.; Naval, PC.; Voit, EO.; Mendoza, E. MADMan: a benchmarking framework for parameter estimation in biochemical systems theory models. International Conference on Molecular Systems Biology 2008 (ICMSB08); Manila, Philippines. 2008. p. 10-13.

239. Sekiguchi T, Okamoto M. WinBEST-KIT: Windows-based Biochemical Reaction Simulator for Metabolic Pathways. J Bioinform Comput Biol 2006;4:621–638. [PubMed: 16960966]

240. Cadlive, CADLIVE (Computer-Aided Design of LIVing systEms). www.cadlive.jp (2009).

241. Almeida, JS. Bioinformatics Station. 2008. http://bioinformaticstation.org

242. Voit EO. The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions. Biometr J 1992;34:855–878.

243. Voit EO, Yu S. The S-distribution: Approximation of discrete distributions. Biometr J 1994;36:205–219.

244. Yu, SS.; Voit, EO. A graphical classification of survival distributions. In: Jewell, NP.; Kimber, AC.; Lee, M-LT.; Whitmore, GA., editors. Lifetime Data: Models in Reliability and Survival Analysis. Kluwer Academic Publishers; Dordrecht: 1996. p. 385-392.

245. Sorribas A, March J, Voit EO. Estimating age-related trends in cross-sectional studies using S-distributions. Stat Med 2000;19:697–713. [PubMed: 10700740]

246. Voit EO. Dynamic trends in distributions. Biometr J 1996;38:587–603.

247. Voit EO, Sorribas A. Computer modeling of dynamically changing distributions of random variables. Math Comput Modelling 2000;31:217–225.

248. Marin-Sanguino A, Voit EO, Gonzalez-Alcon C, Torres NV. Optimization of biotechnological systems through geometric programming. Theor Biol Med Model 2007;4:38. [PubMed: 17897440]

249. Neves AR, Ramos A, Nunes MC, Kleerebezem M, Hugenholtz J, de Vos WM, Almeida J, Santos H. In vivo nuclear magnetic resonance studies of glycolytic kinetics in Lactococcus lactis. Biotechnol Bioeng 1999;64:200–212. [PubMed: 10397856]

250. Neves AR, Ramos A, Shearman C, Gasson MJ, Almeida JS, Santos H. Metabolic characterization of Lactococcus lactis deficient in lactate dehydrogenase using in vivo 13C-NMR. Eur J Biochem 2000;267:3859–3868. [PubMed: 10849005]

251. Vera J, de Atauri P, Cascante M, Torres NV. Multicriteria optimization of biochemical systems by linear programming: application to production of ethanol by *Saccharomyces cerevisiae*. Biotechnol Bioeng 2003;83:335–343. [PubMed: 12783489]

252. Sutton MD, Smith BT, Godoy VG, Walker GC. The SOS response: recent insights into umuDC-dependent mutagenesis and DNA damage tolerance. Annu Rev Genet 2000;34:479–497. [PubMed: 11092836]

253. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 2005;33:D562–566. [PubMed: 15608262]

254. Kuper C, Jung K. CadC-mediated activation of the cadBA promoter in *Escherichia coli*. J Mol Microbiol Biotechnol 2005;10:26–39. [PubMed: 16491024]

255. Xiu ZL, Chang ZY, Zeng AP. Nonlinear dynamics of regulation of bacterial trp operon: Model analysis of integrated effects of repression, feedback inhibition, and attenuation. Biotechnol Prog 2002;18:686–693. [PubMed: 12153299]

256. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998;2:65–73. [PubMed: 9702192]

257. Neves AR, Pool WA, Kok J, Kuipers OP, Santos H. Overview on sugar metabolism and its control in Lactococcus lactis - the input from in vivo NMR. FEMS Microbiol Rev 2005;29:531–554. [PubMed: 15939503]

258. Wang FS, Su TL, Jang HJ. Hybrid differential evolution for problems of kinetic parameter estimation and dynamic optimization of an ethanol fermentation process. Industrial & Engineering Chemistry Research 2001;40:2876–2885.

259. Ingalls BP. Autonomously oscillating biochemical systems: Parametric sensitivity of extrema and period. Syst Biol (Stevenage) 2004;1:62–70. [PubMed: 17052116]

260. Huang, WH.; Yuh, CH.; Wang, FS. Reverse engineering for embryonic gene regulatory network in zebrafish via evolutionary optimization with data collocation. 7th International Conference on Systems Biology; Yokohama, Japan. 2006.

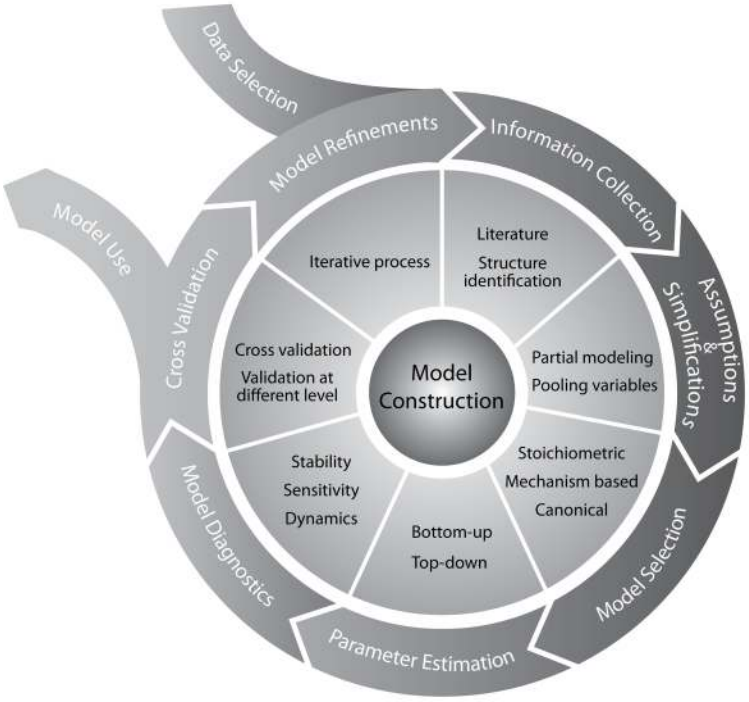

Fig. 1.
Phases of the typical modeling process in biology. See *Text* for details.

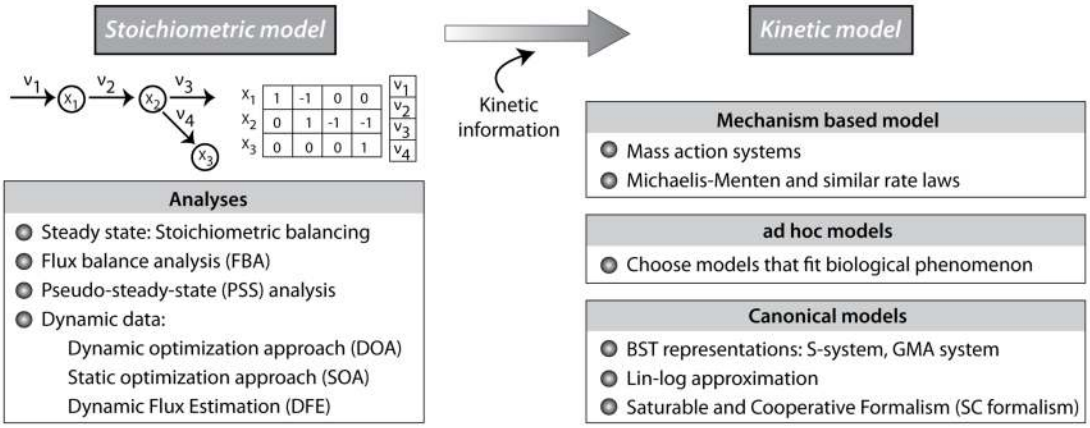

Fig. 2.
Traditional models used in metabolic systems analysis. Stoichiometric models focus on the connectivity of the biological system. By incorporating kinetic information, kinetic models are able to describe the dynamics of the metabolic pathway.

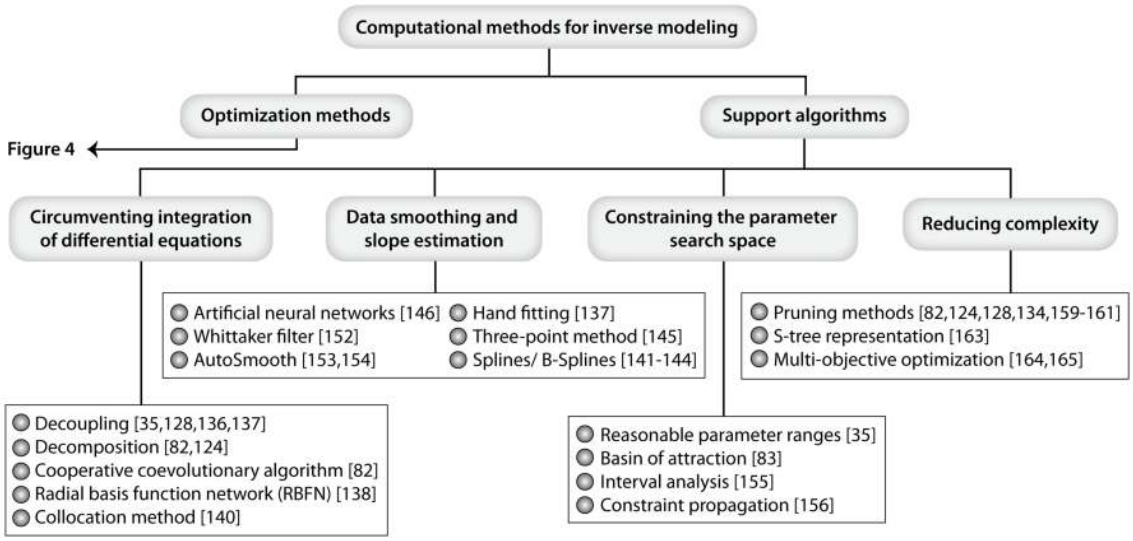

Fig. 3.
Support algorithms for inverse modeling. Some representative references are listed.

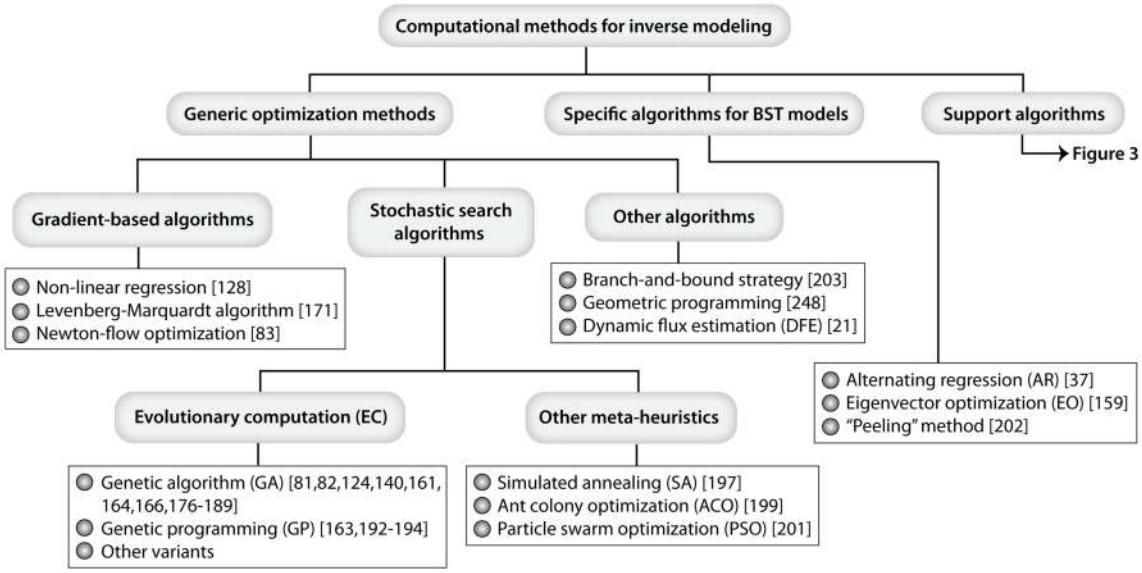

Fig. 4.
Optimization algorithms for inverse modeling. Some representative references are listed.

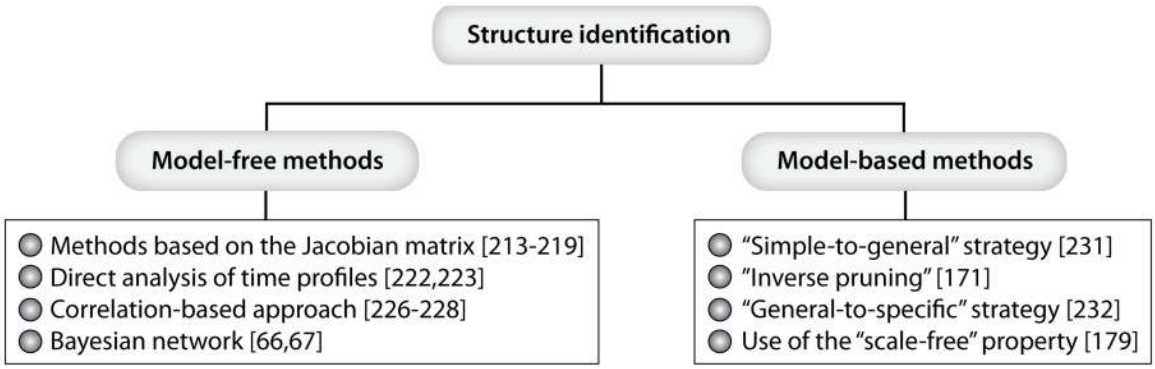

Fig. 5.
Structure identification algorithms for inverse modeling. Some representative references are listed.

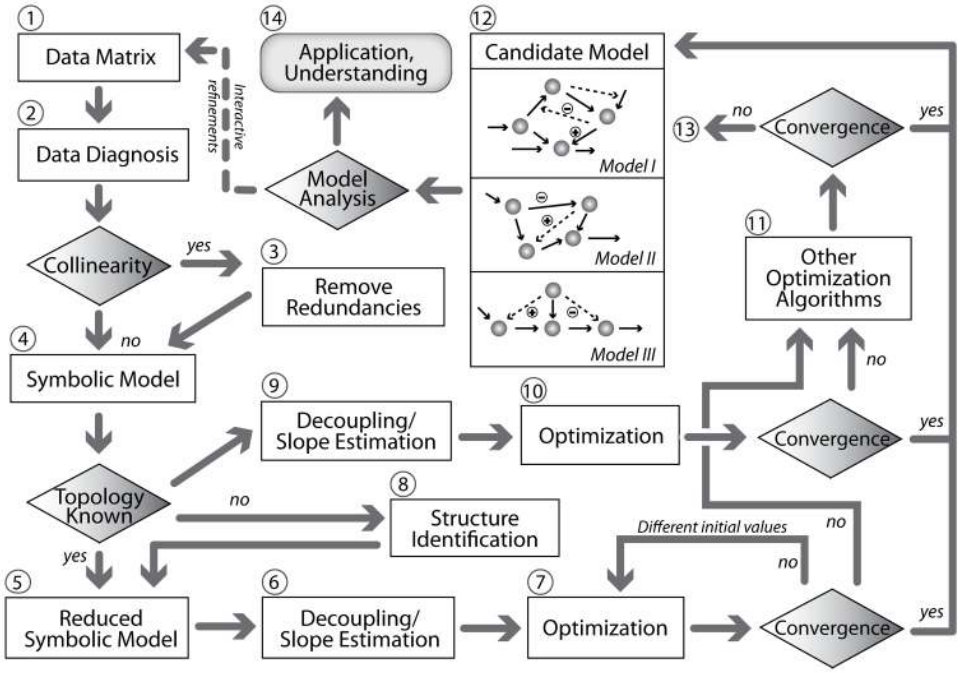

Fig. 6.
Flow diagram of inverse modeling. See *Text* for details.

Table 1
Organization of the Reviewed Material

Section 2: Modeling approaches
2.1 Model requirements
2.2 Stoichiometric pathway models
2.3 Kinetic models of pathway steps
2.3.1 Mechanistically based functions
2.3.2 *Ad hoc* modeling approaches
2.3.3 Canonical models
2.3.4 Dynamic models of gene regulatory networks
Section 3: Kinetic model construction
3.1 Forward or bottom-up modeling
3.2 Model retrieval from steady-state data
3.3 Inverse or top-down modeling
3.4 Challenges of the top-down modeling approach and current solution strategies
3.4.1 Data related issues
3.4.2 Model related issues
3.4.3 Computational issues
3.4.4 Mathematical issues
Section 4: Parameter estimation techniques for top-down modeling approaches
4.1 Methods based on integrating differential equations
4.2 Slope estimation
4.3 Constraining the parameter search space
4.4 Reducing the complexity of the inference task
4.5 Algorithms for determining optimal parameter estimates
4.5.1 Gradient-based algorithms
4.5.2 Stochastic search algorithms
4.5.3 Other algorithms
Section 5: Inference of network structure
5.1 Model-free structure identification approaches
5.1.1 Methods based on the Jacobian matrix
5.1.2 Direct observation
5.1.3 Correlation-based approach
5.1.4 Bayesian network approach
5.2 Model-based structure identification methods
5.2.1 ‘Simple-to-general’ and ‘general-to-specific’ modeling
5.2.2 Use of time series data
Section 6: Toward a streamlined “work-flow” for inverse modeling
6.1 Benchmarking framework
6.2 Work-flow strategy
6.2.1 Goals
6.2.2 Flow diagram of inverse modeling strategy
Section 7: Open issues

Table 2

Challenges and current solutions of parameter estimation and structure identification tasks in inverse modeling.

Inverse Modeling Approach	Tasks	Challenges	Solutions
	Parameter Estimation ⇕ Structure Identification	**Data** • Overly noisy data • Missing data points • Uncertainties about the measurements • Ill-posed data matrix • Non-informative data profile	• Check data consistency • Data diagnoses (*e.g.* collinearity) • Data preprocessing (*e.g.* pooling variables) • Concept map modeling
		Model • Model selection criteria - Dynamic flexibility - Mathematical approximation - Mathematical tractability - Interpretability of results	• BST models: S-system, GMA • Lin-log approximation • Saturable and Cooperative Formalism (SC formalism) • Determination of model suitability
		Computation • Computational capacity • Slow convergence • Lacking convergence or convergence to local minima • Computational cost for integration of differential equations	• Optimization methods • Supporting algorithms - complexity reduction - Avoiding ODE integration - Data smoothing and slope estimation - Parameter search space constraints
		Math • Distinctly different yet equivalent solutions • Non-equivalent solutions with similar error • Error compensation	• Estimation of fluxes • Data covering wide ranges of variation • Multiple datasets • Additional information about some of the parameter values
		Topology (structure identification)	• Model-free, coarse methods • Model based methods

Table 3

Timeline of representative algorithms for inverse problems in BST models.

Authors	Year	Main Methods and Features	Model	Target Networks	
				Artificial*	Actual**
Voit and Savageau [136]	1982	• Decoupling	S-system		(a)
Voit [35]	2000	• Review of various bottom-up and top- down methods	S-system		
			GMA		
Seatzu [144]	2000	• Smoothing (B-splines)	S-system		(b)
Maki *et al.* [123]	2002	• "Step-by-step" strategy	S-system		(c)
Kikuchi *et al.* [134]	2003	• Simple genetic algorithm (SGA) • Penalty term in the objective function	S-system	(A)	
Kimura *et al.* [124]	2004	• Decomposition method • Numerical integration with local linear regression	S-system	(A) (B)	
Voit and Almeida [128]	2004	• Decoupling • ANN smoothing and slope approximation	S-system	(C)	
Kimura *et al.* [82]	2005	• Decomposition • Cooperative coevolution algorithm	S-system	(A) (B)	(d)
Lall and Voit [202]	2005	• "Peeling" technique	S-system		(e)
Tsai and Wang [140]	2005	• Modified collocation method	S-system	(A) (D)	
Cho *et al.* [163]	2006	• S-tree based genetic programming (GP)	S-system	(A)	(f) (g)
Chou *et al.* [37]	2006	• Alternating regression (AR)	S-system	(A) (E)	
Daisuke and Horton [179]	2006	• Distributed genetic algorithm (DGA) • Use of scale-free property	S-system	(A)	(h)
Kim *et al.* [194]	2006	• Genetic programming to estimate slopes and avoid numerical integration	S-system	(E)	
Marino and Voit [171]	2006	• Gradual increase in model complexity	S-system	(C)	
Naval *et al.* [201]	2006	• Particle swarm optimization (PSO)	S-system	(C)	
			GMA		(i)
Polisetty *et al.* [203]	2006	• Branch-and-reduce strategy	GMA	(F)	(i)
Tucker and Moulton[155]	2006	• Interval analysis	S-system	(A) (E) (G)	
Gonzalez *et al.* [197]	2007	• Simulated annealing (SA)	S-system	(C)	(j)
Kutalik *et al.* [83]	2007	• Newton-flow method	S-system	(B) (E) (H) (I)	
Marin-Sanguino *et al.* [248]	2007	• GMA optimizer • Geometric programming	GMA		(i) (k)
Noman and Iba [189]	2007	• Information criteria-based fitness evaluation	S-system	(A) (J)	(l)

Authors	Year	Main Methods and Features	Model	Target Networks	
				Artificial*	Actual**
		• Differential evolution (DE) along with local search heuristics			
Tucker *et al*. [156]	2007	• Constraint propagation	S-system	(E)	
			GMA	(K)	
Goel et al. [21]	2008	• Dynamic flux estimation (DFE)	GMA		(m)
Liu and Wang [164]	2008	• Multi-objective optimization	S-system	(A) (B)	(n) (o) (p)
Vilela *et al*. [159]	2008	• Eigenvector optimization (EO)	S-system	(A) (E) (H) (L)	
Zuñiga *et al*. [199]	2008	• Ant colony optimization (ACO) • Enhanced aggregation pheromone system (eAPS)	S-system		

*The artificial target networks used in the representative algorithms are: (A) Five-variable gene regulatory network [72]; (B) Thirty-variable system [81]; (C) Five-variable didactic system (four dependent variables and one independent variable) [128]; (D) Three-variable cascaded system [140]; (E) Four-variable didactic system (similar pathway as model (C) but without independent variables) [37]; (F) Four-variable branched pathway with several feedback inhibitions (three dependent variables and one independent variable) [35]; (G) Three-variable cascaded pathway [35]; (H) Two-variable system [83]; (I) Seven-variable system [83]; (J) Twenty-variable system [189]; (K) Three-variable branched pathway with several feedback inhibition signals (similar pathway as model (F) but without independent variables) [35]; (L) Ten-variable system ([159]).

**The real networks used in the representative algorithms are: (a) Four-variable model of ethanol production by yeast [136]; (b) Five-variable forest growth model (four dependent variables and one independent variable) [31]; (c) Gene expression profiles during neural differentiation of P19 EC cells measured with mouse cDNA microarrays representing 15,000 genes [123]; (d) cDNA microarray data of *Thermus thermophilus* HB8 strains [82]; (e) NMR data from the *L. lactis* glycolysis pathway (model described in [202]; experimental data from [115,249,250]); (f) Anaerobic fermentation pathway in *Saccharomyces cerevisiae* (five dependent variables and eight independent variables) [251]; (g) SOS DNA repair system in *E. coli* [252]; (h) Gene expression profiles of mice (data selected from GDS404 in NCBI [253]) [179]; (i) Anaerobic fermentation pathway in *Saccharomyces cerevisiae* (same pathway as in model (f) but GMA model) [95]; (j) cadBA in *E. coli* [254]; (k) Tryptophan operon model in *E. coli* [255]; (l) Yeast cell-cycle microarray data [256]; (m) NMR data from the *L. lactis* glycolysis pathway [257] (same pathway as pathway (e) but GMA model [21]); (n) Kinetic model of ethanol fermentation [258]; (o) Circadian oscillations of period proteins in drosophila [259]; (p) Embryonic gene regulatory network in zebrafish [260].